

FIG. 1

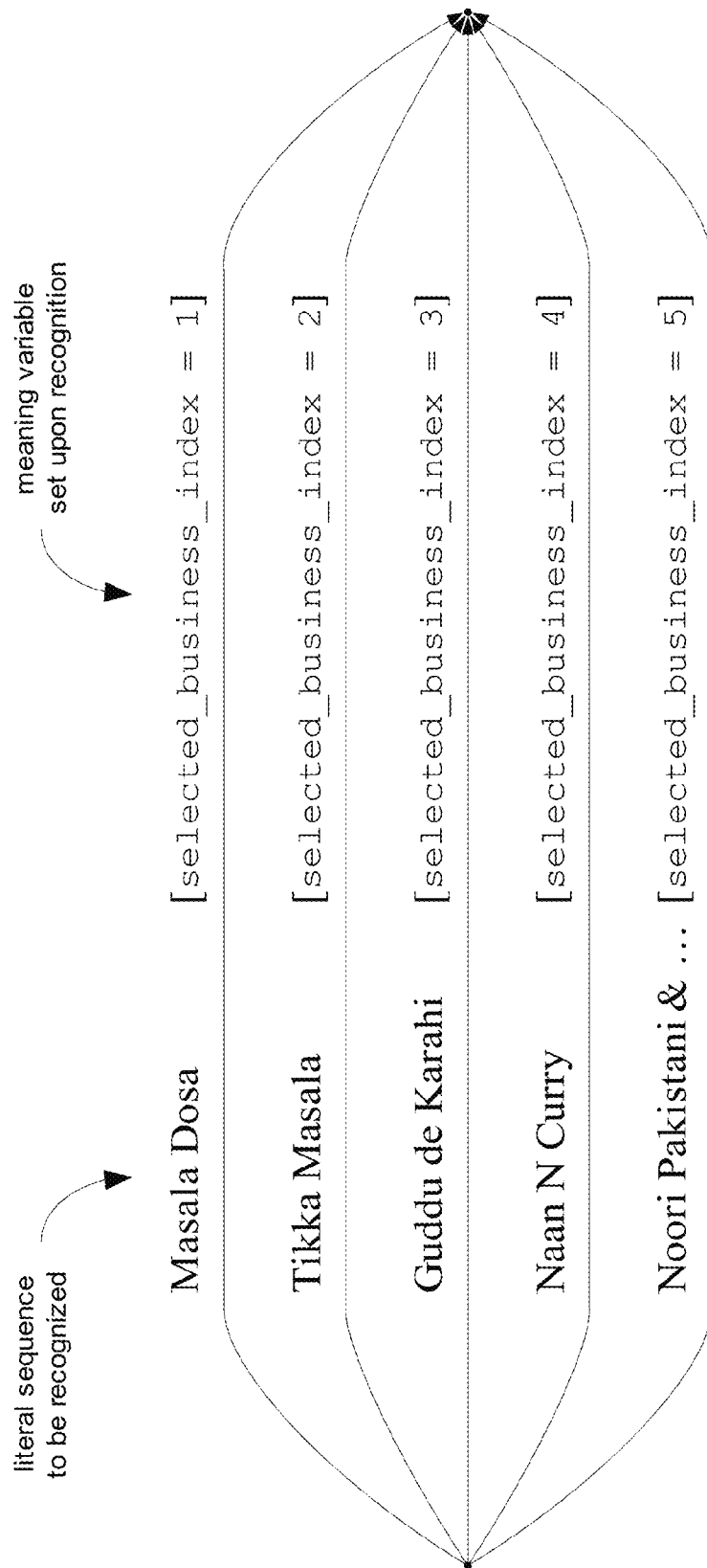


FIG. 2

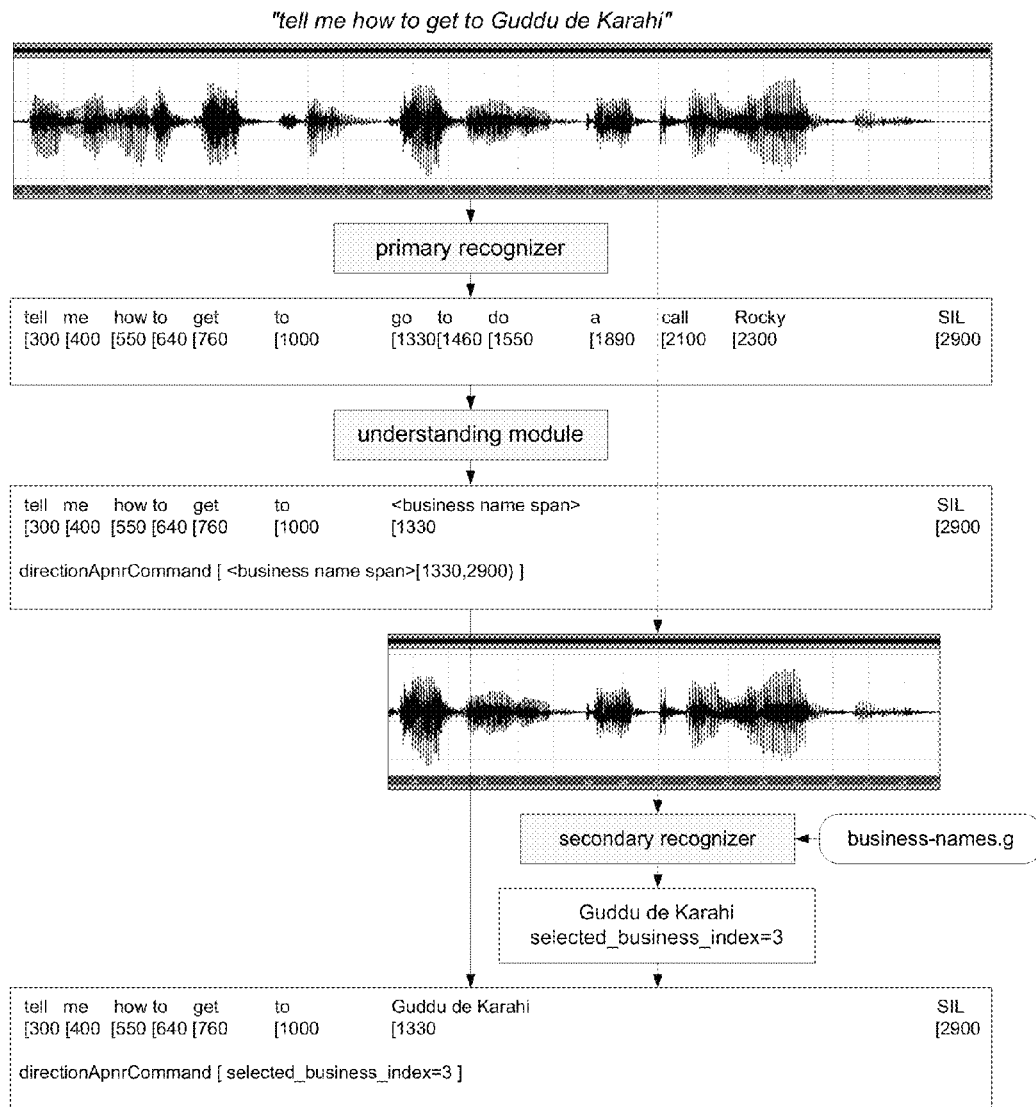


FIG. 3

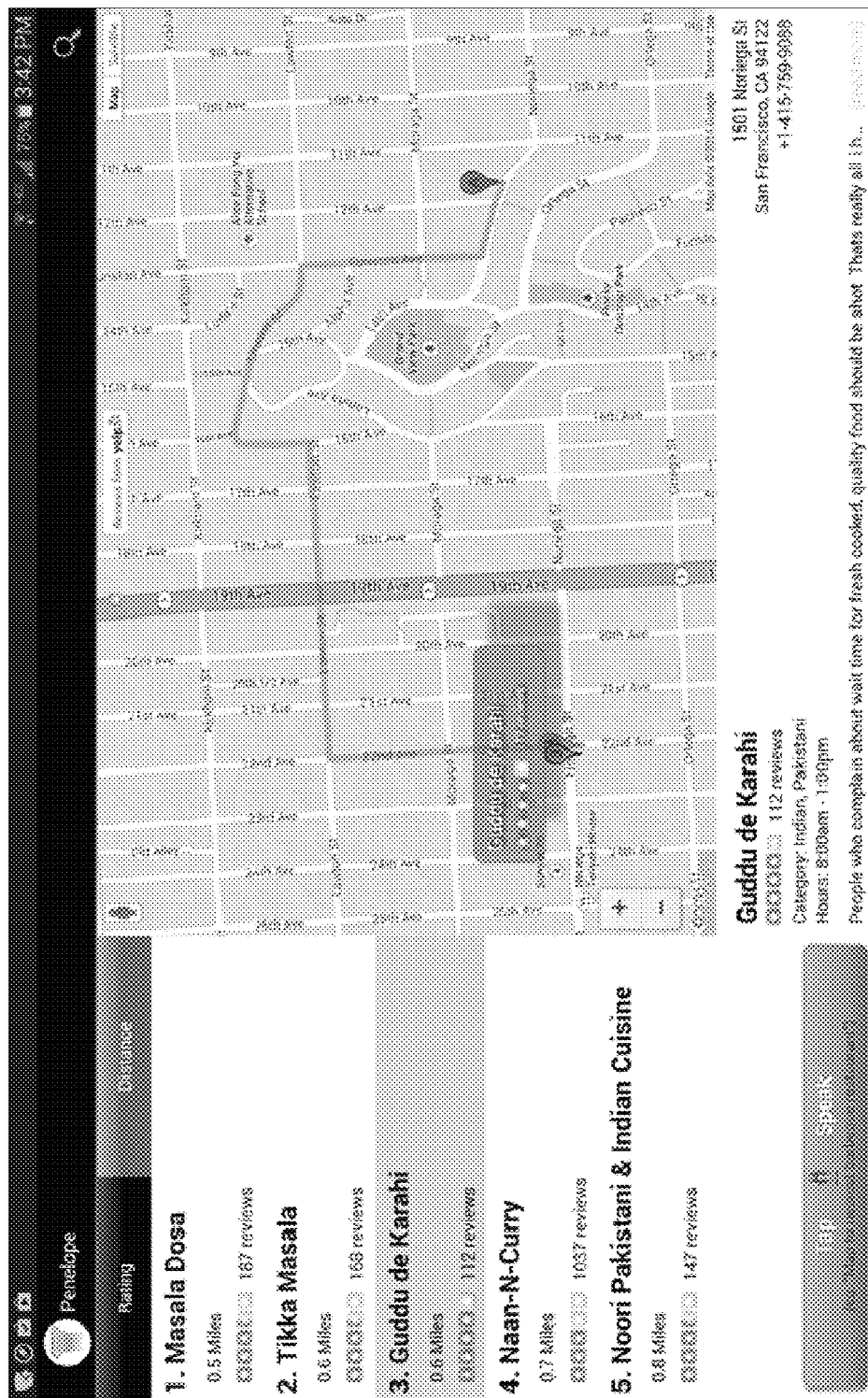


FIG. 4

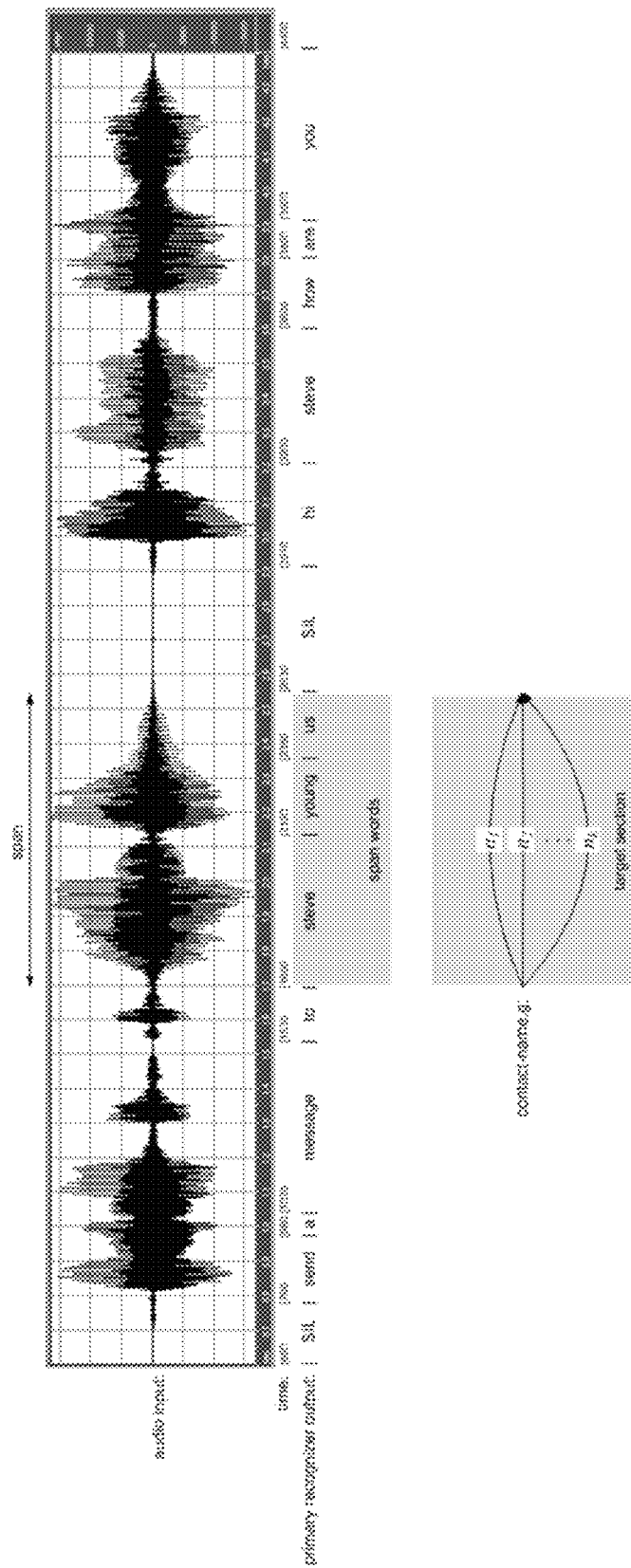


FIG. 5

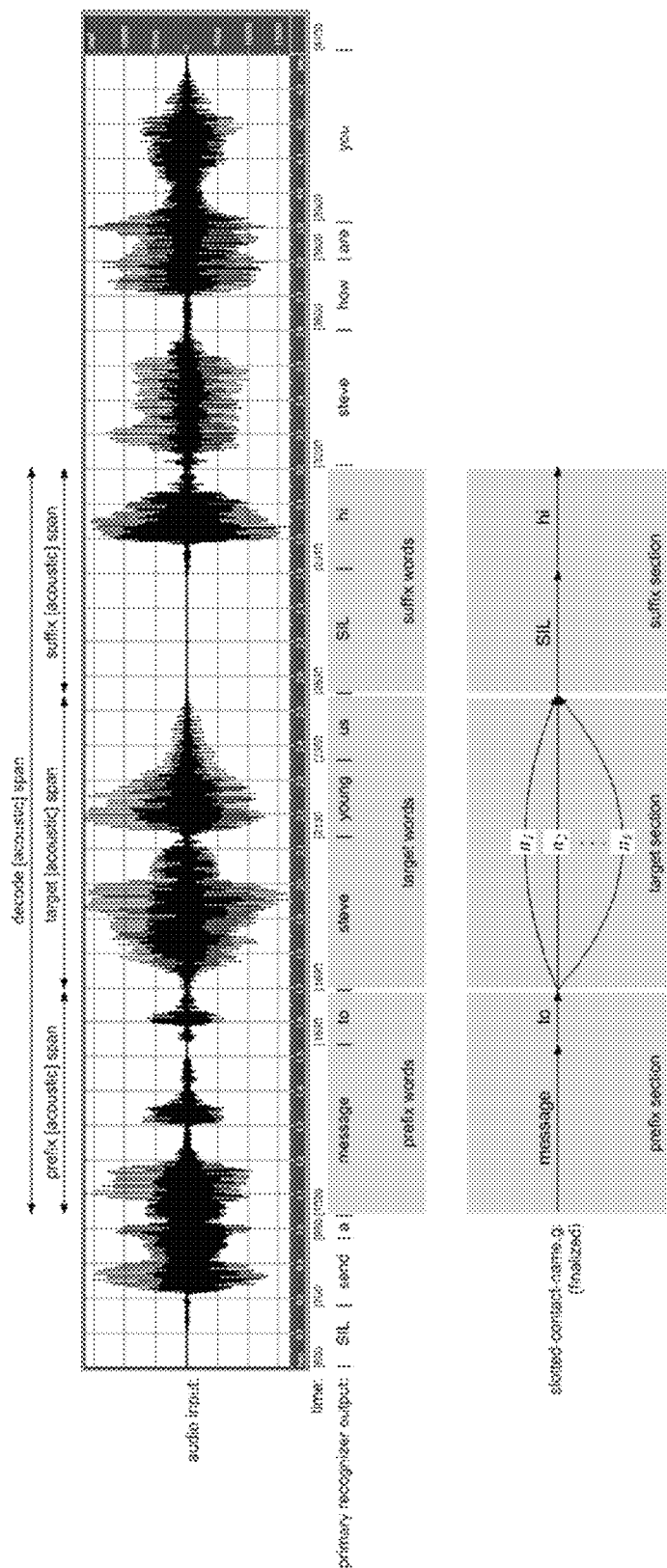


FIG. 6

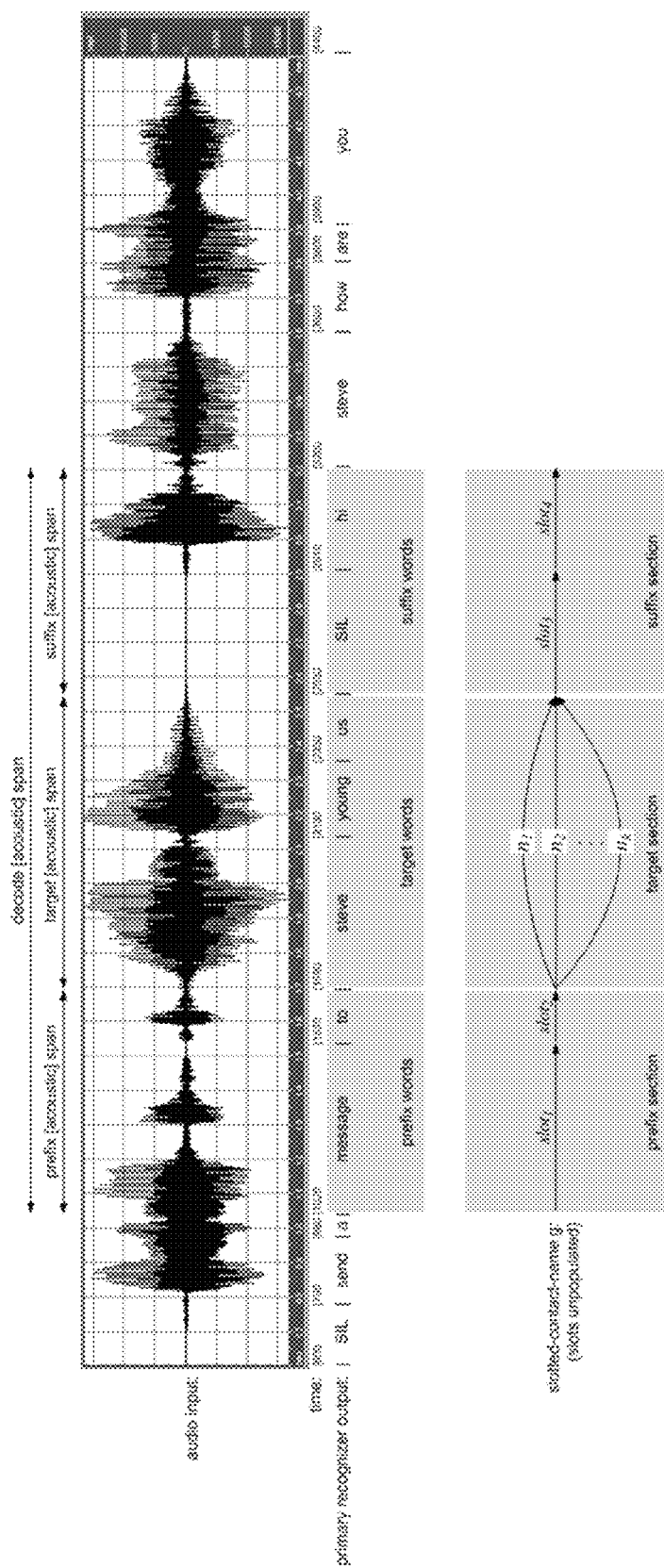
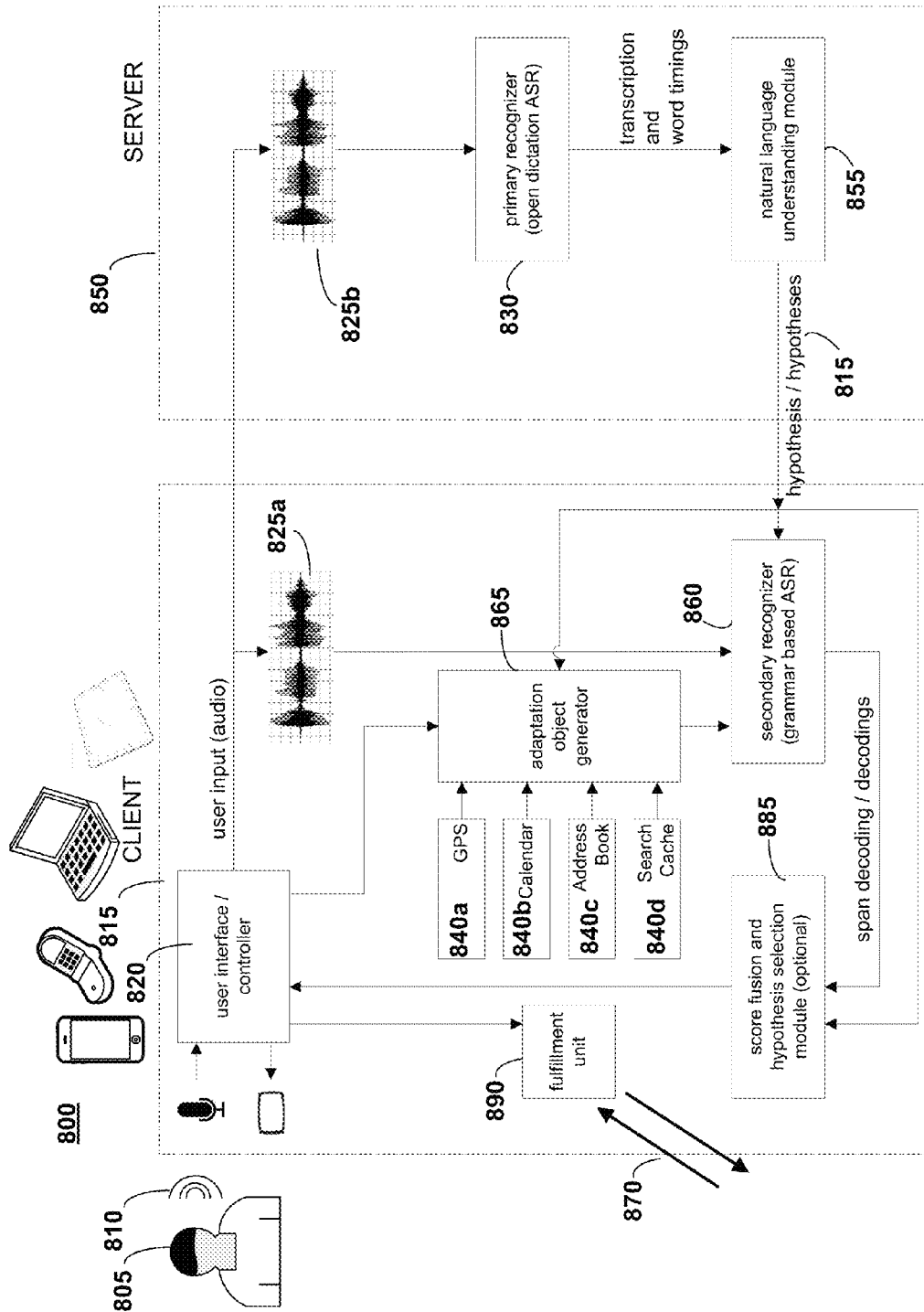
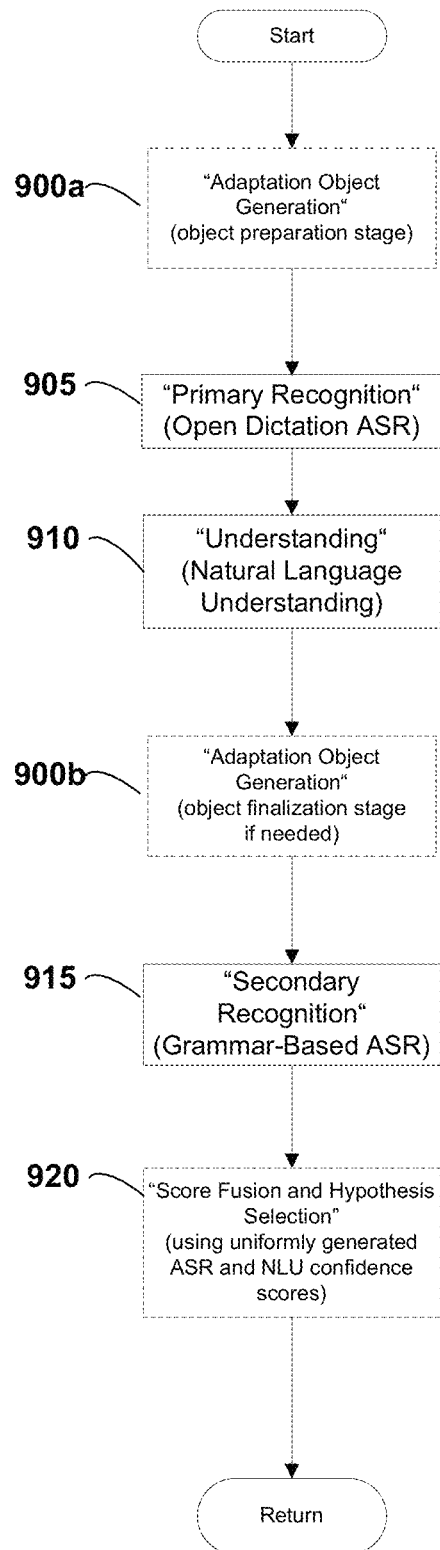


FIG. 7



**FIG. 9**

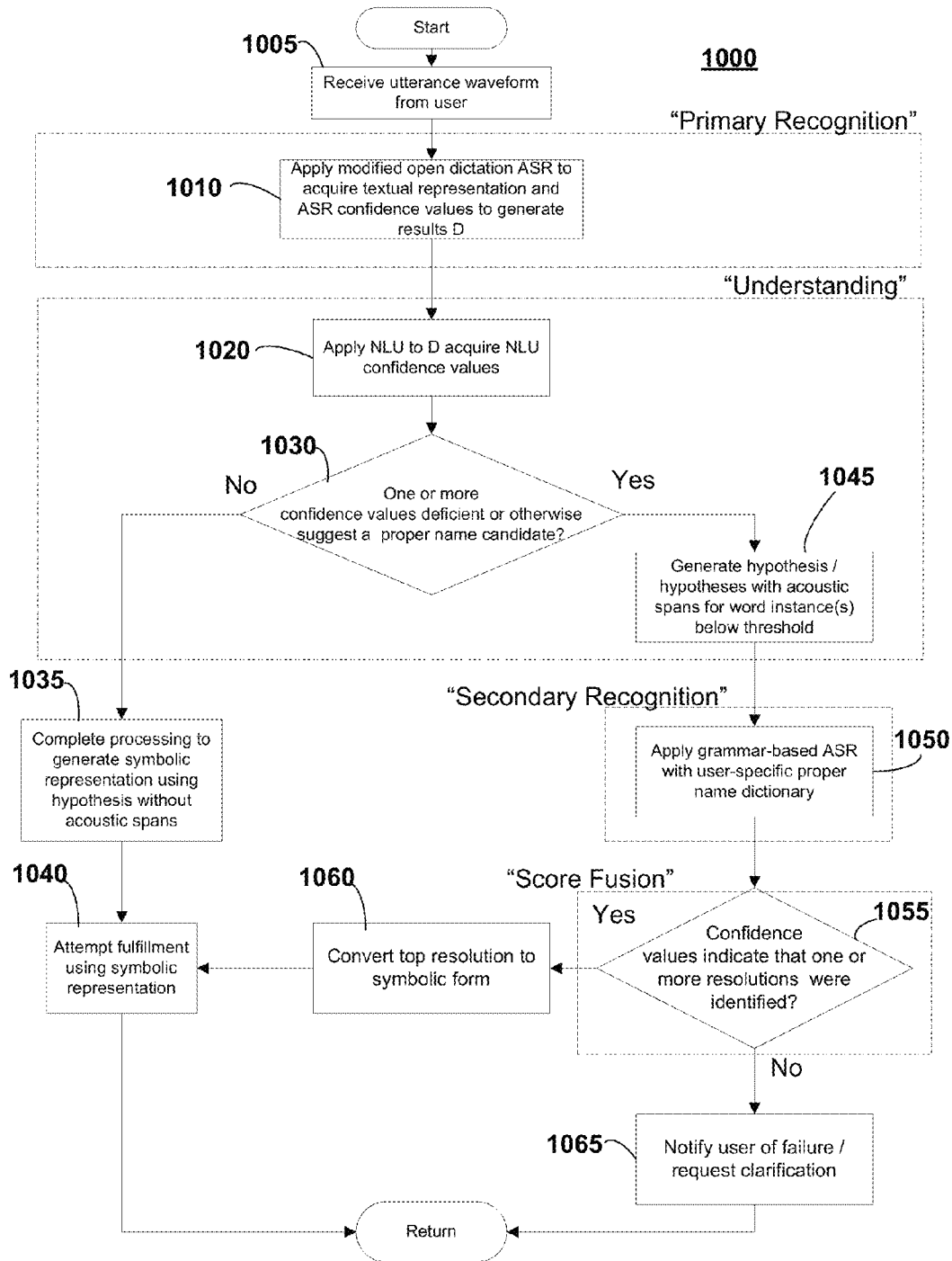


FIG. 10

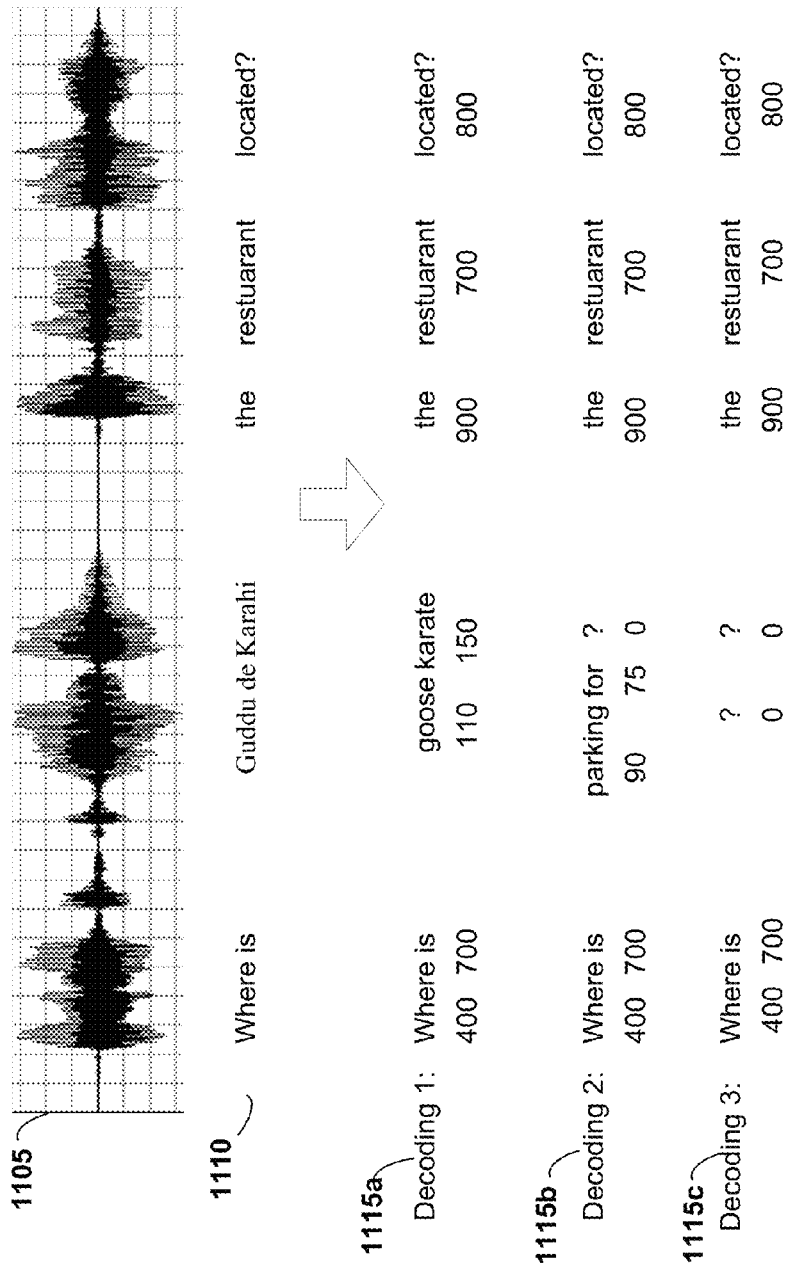


FIG. 11

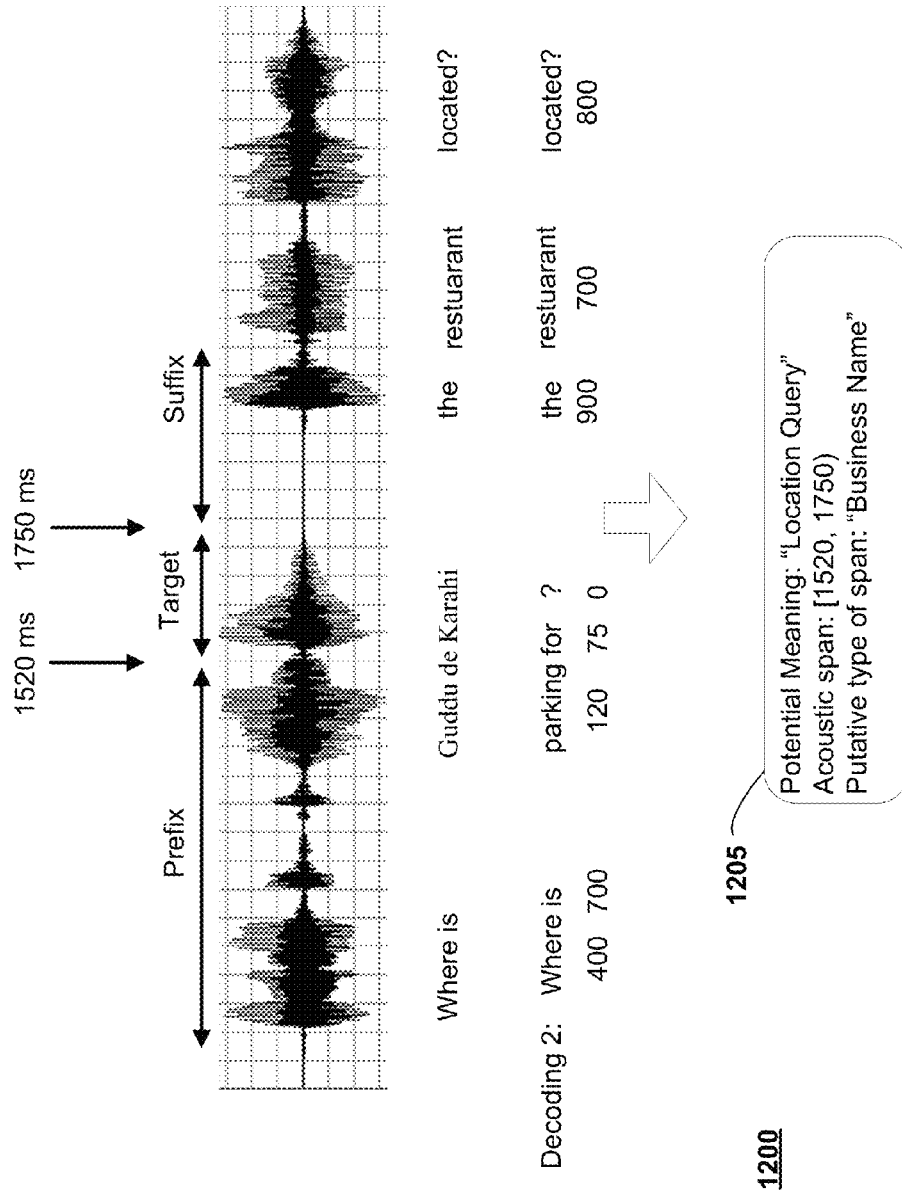


FIG. 12

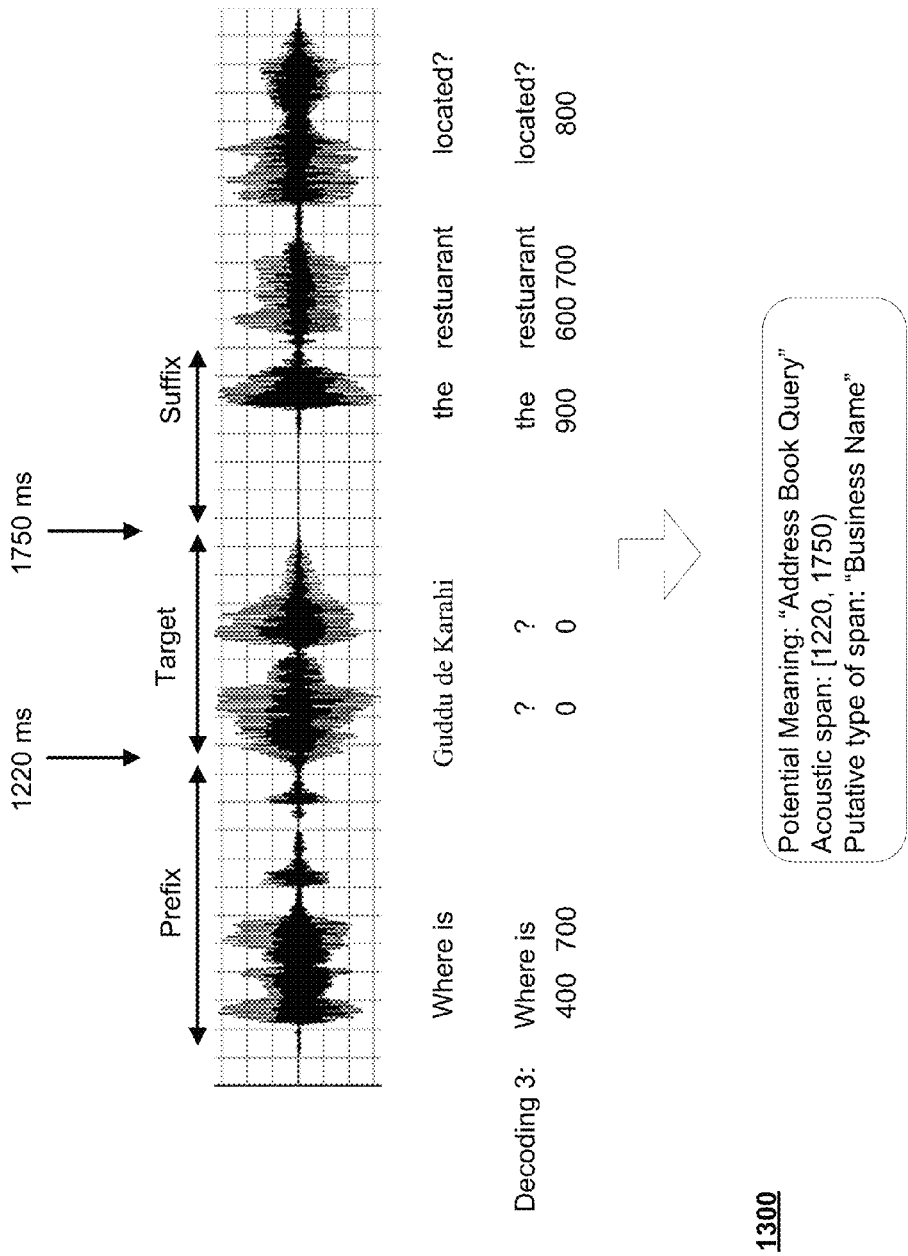
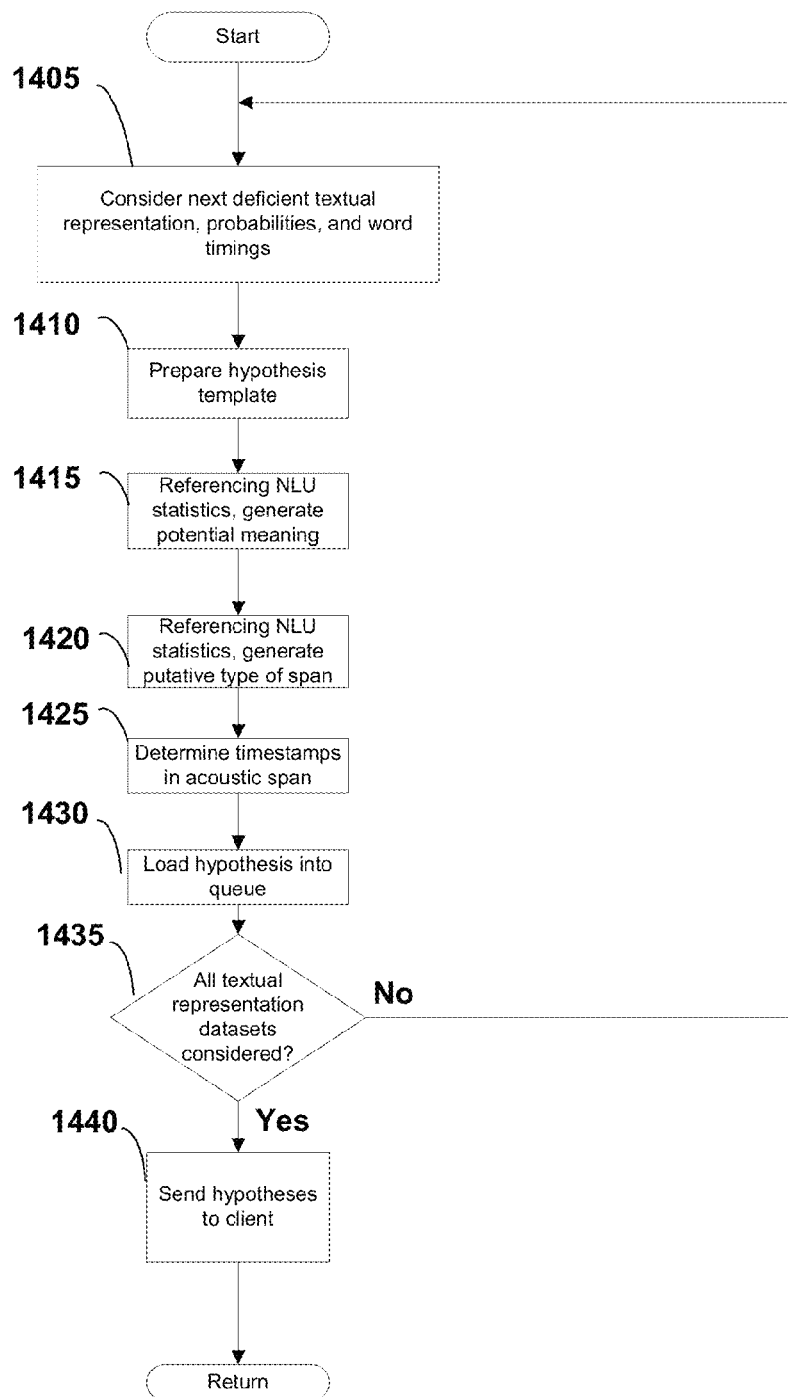
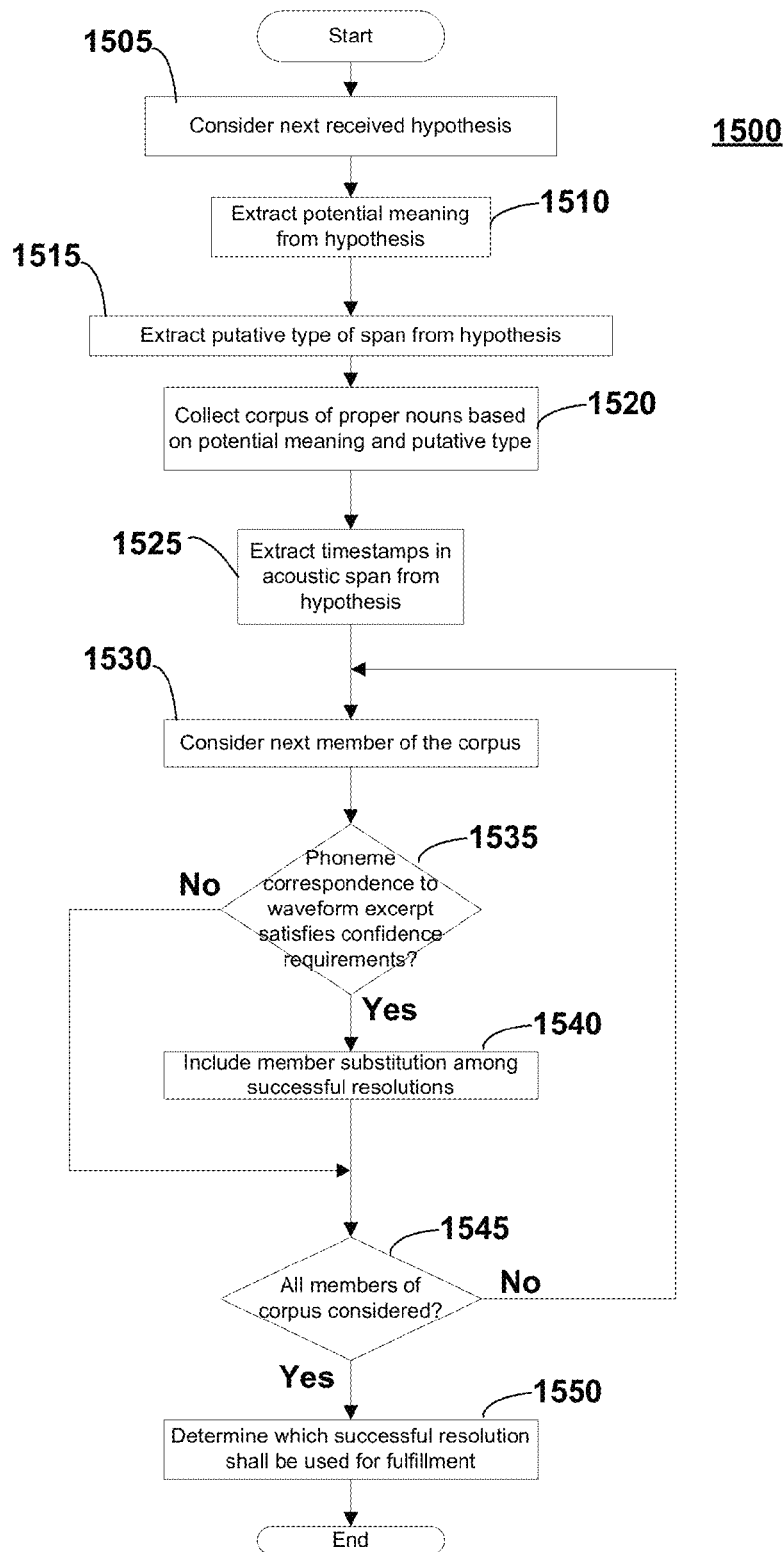


FIG. 13

1400**FIG. 14**

**FIG. 15**

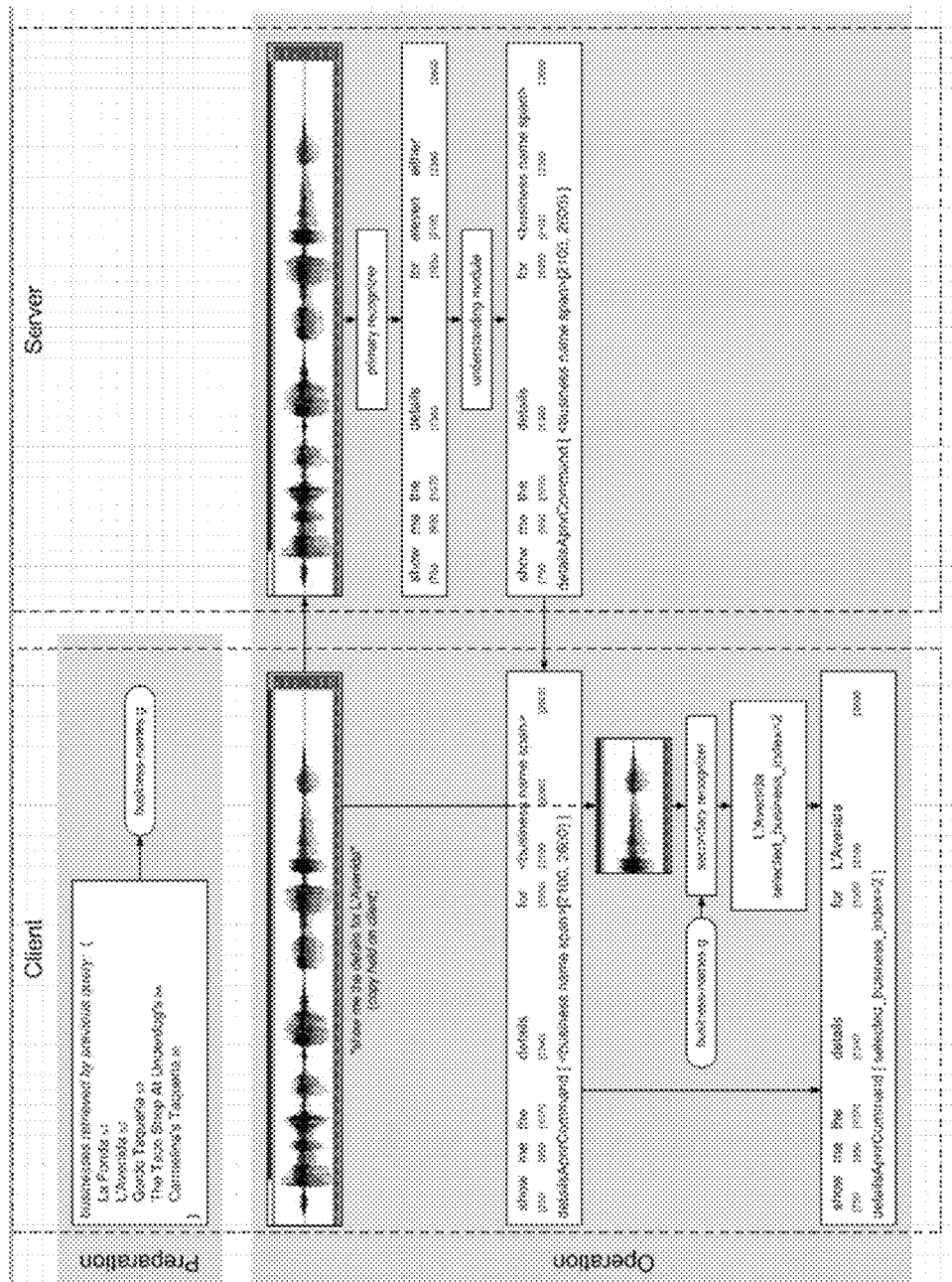


FIG. 16

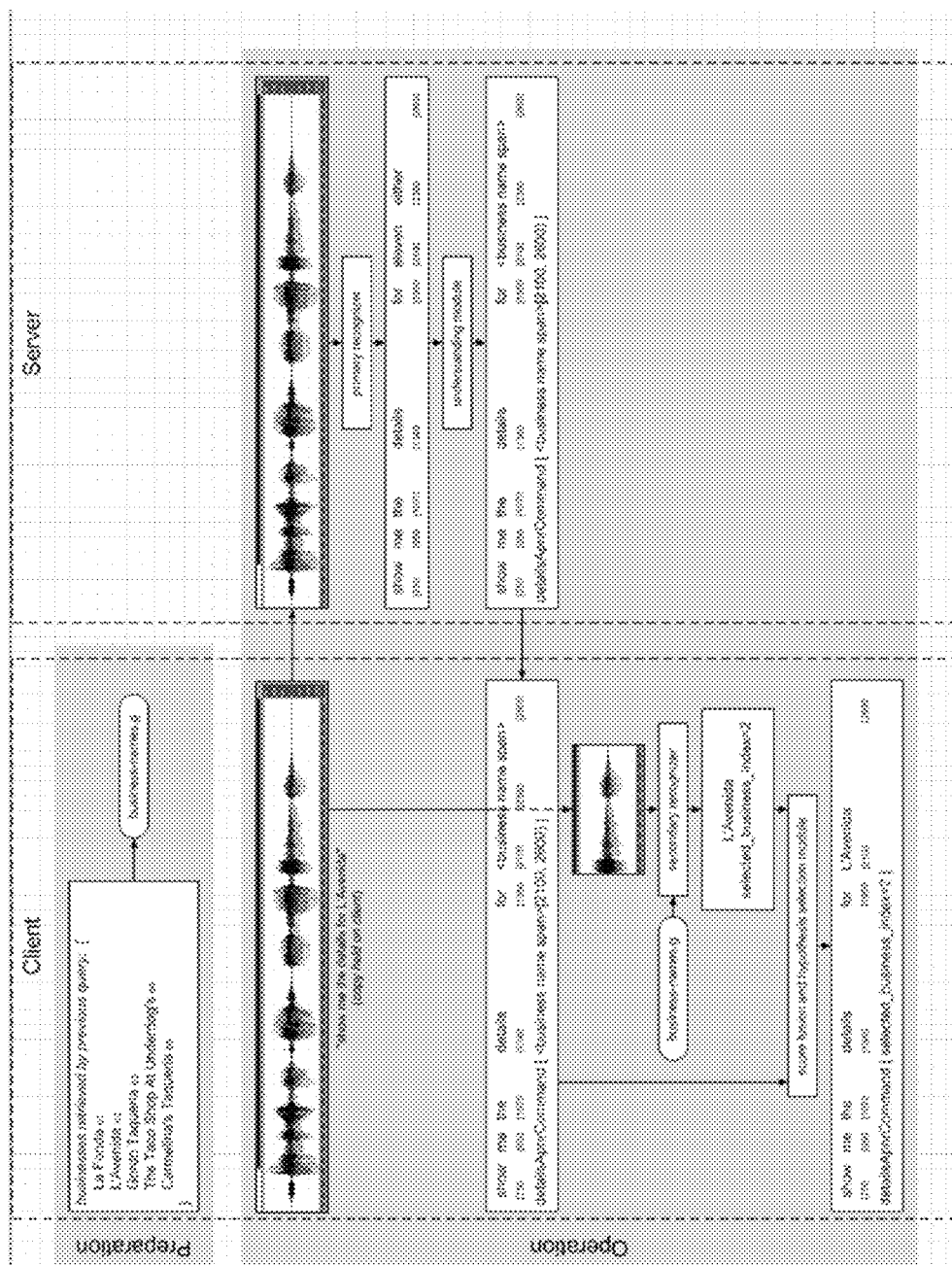


FIG. 17

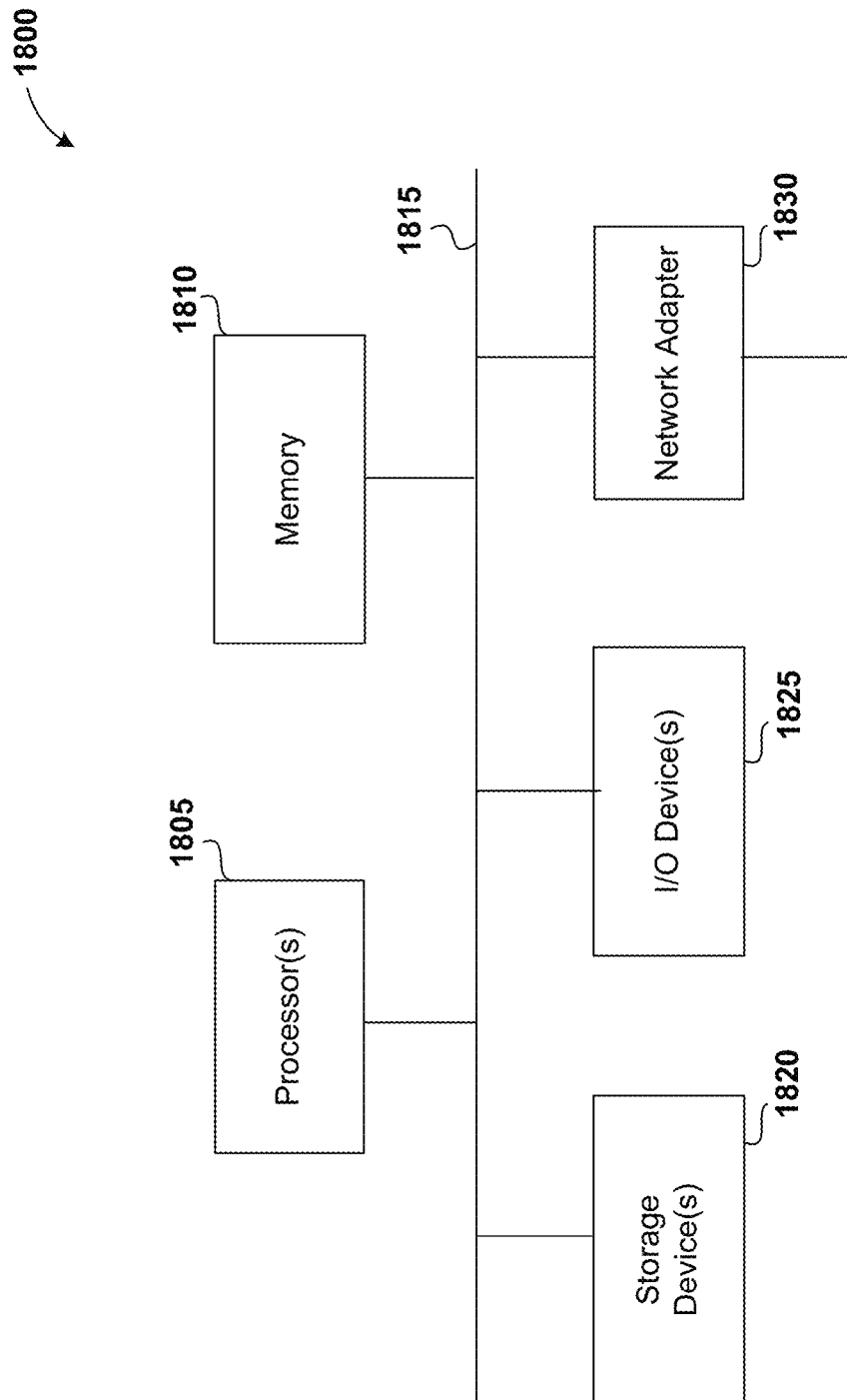


FIG. 18

SYSTEMS AND METHODS FOR ADAPTIVE PROPER NAME ENTITY RECOGNITION AND UNDERSTANDING

CROSS-REFERENCE TO RELATED APPLICATIONS

This application is entitled to the benefit of and claims priority to U.S. Provisional Patent Application No. 61/828, 919, entitled “Adaptive Proper Name Recognition and Understanding” filed May 30, 2013, the contents of which are incorporated herein by reference in their entirety for all purposes.

FIELD

Various of the disclosed embodiments relate to systems and methods for automatic recognition and understanding of fluent, natural human speech, notably speech that may include proper name entities, as discussed herein.

BACKGROUND AND PROBLEM SOLVED

Automatic speech recognition (ASR) technology and natural language understanding (NLU) technology have advanced significantly in the past decade, ushering in the era of the spoken language interface. For example, the “Siri®” system, which allows users to speak a multitude of questions and commands to the “iPhone®” cellular telephone and Google’s similar “Google Voice™” service, have gained mass-market acceptance.

While such products are remarkably successful at recognizing generic requests like “set a reminder for Dad’s birthday on December 1st” or “what does my calendar look like for today,” they can be foiled by utterances that contain proper names, especially uncommon ones. Commands like “set my destination to Barbagelata Real Estate,” “tell me how to get to Guddu de Karahi,” or “give me the details for Narayanaswamy Harish, DVM”—all of which are reasonable requests, within appropriate contexts—often yield results that are incorrect if not outright comical.

Accordingly, there is a need for systems providing more accurate recognition of proper names.

BRIEF DESCRIPTION OF THE DRAWINGS

The techniques introduced here may be better understood by referring to the following Detailed Description in conjunction with the accompanying drawings, in which like reference numerals indicate identical or functionally similar elements:

FIG. 1 is a screenshot of an example graphical user interface in a personal assistant application implementing various features of some embodiments.

FIG. 2 is a graphical depiction of a grammar generated using a request in the example of FIG. 1 in some embodiments.

FIG. 3 is an example processing diagram depicting the processing operations of an embodiment as applied to an example word sentence.

FIG. 4 is a screenshot of an example graphical user interface depicting the results following processing in an example system as may occur in some embodiments.

FIG. 5 is an example breakdown of an utterance waveform as may occur in some embodiments.

FIG. 6 is an example breakdown of an utterance waveform as may occur in some embodiments.

FIG. 7 is an example breakdown of an utterance waveform as may occur in some embodiments.

FIG. 8 is a block diagram depicting various components in an example speech processing system having server and client proper name resolution modules as may occur in some embodiments.

FIG. 9 is a flow diagram depicting the proper name recognition process at a high level for various embodiments using automatic speech recognition (ASR) and natural language understanding (NLU) components.

FIG. 10 is a flow diagram depicting various steps in a proper name recognition process as may occur in some embodiments.

FIG. 11 is an example hypothesis corpus as may be generated in some embodiments.

FIG. 12 is an example of a first hypothesis breakdown based upon the example of FIG. 11 as may occur in some embodiments.

FIG. 13 is an example of a second hypothesis breakdown based upon the example of FIG. 11 as may occur in some embodiments.

FIG. 14 is a flow diagram depicting various steps in a server-side process for proper name recognition as may occur in some embodiments.

FIG. 15 is a flow diagram depicting various steps in a client-side process for proper name recognition as may occur in some embodiments.

FIG. 16 and FIG. 17 are further processing examples without and with score fusion respectively as may occur in some embodiments.

FIG. 18 is a block diagram of a computer system as may be used to implement features of some of the embodiments.

The headings provided herein are for convenience only and do not necessarily affect the scope or meaning of the claimed embodiments. Further, the drawings have not necessarily been drawn to scale. For example, the dimensions of some of the elements in the figures may be expanded or reduced to help improve the understanding of the embodiments. Similarly, some components and/or operations may be separated into different blocks or combined into a single block for the purposes of discussion of some of the embodiments. Moreover, while the various embodiments are amenable to various modifications and alternative forms, specific embodiments have been shown by way of example in the drawings and are described in detail below. The intention, however, is not to limit the particular embodiments described. On the contrary, the embodiments are intended to cover all modifications, equivalents, and alternatives falling within the scope of the disclosed embodiments as defined by the appended claims.

GLOSSARY

The following glossary is provided as a convenience to the reader, collecting in one place the acronyms, abbreviations, symbols and specialized terminology used throughout this specification.

An “acoustic prefix” as referenced herein is one or more words, as decoded in the primary recognition step, that precede a target span. This may also be called the “left acoustic context.”

An “acoustic span” is a portion of an audio waveform.

An “acoustic suffix” is one or more words, as decoded in the primary recognition step, that follow a target span. This may also be called the “right acoustic context.”

An “adaptation object” is computer-stored information that enables adaptation (in some embodiments, very rapid

adaptation) of a secondary recognizer to a specified collection of recognizable words and word sequences. For grammar-based ASR systems, this is a grammar, which may be in compiled or finalized form.

An “adaptation object generation module” creates adaptation objects. It may accept as input words or word sequences, some of which may be completely novel, and specifications of allowed ways of assembling the given words or word sequences.

An “adaptation object generator” is the same as an “adaptation object generation module.”

An “adaptation object generation step” is a step in the operation of some embodiments, which may comprise the use of an adaptation object generation module, operating upon appropriate inputs, to create an adaptation object. This process may be divided into two stages, respectively object preparation and object finalization. If the secondary recognizer uses grammar-based ASR technology, “object preparation” may comprise grammar compilation, and “object finalization” may comprise population of grammar slots.

An “aggregate word” is a notional “word,” with very many pronunciations, that stands for an entire collection of proper names. This may be the same as a “placeholder” or “placeholder word.”

“ARPAbet” refers to a phonetic alphabet for the English language. See <http://en.wikipedia.org/wiki/Arpabet>

“ASR” refers to automatic speech recognition: the automatic conversion of spoken language into text.

An “ASR confidence score” refers to a numerical score that reflects the strength of evidence for a particular transcription of a given audio signal.

A “baseform” refers to a triple that associates: (1) a word as a lexical object (that is, a sequence of letters as a word is typically spelled); (2) an index that can be used to distinguish many baseforms for the same word from one another; and (3) a pronunciation for the word, comprising a sequence of phonemes. A given word may have several associated baseforms, distinguished by their pronunciation. For instance, here are the baseforms for the word “tomato”, which as memorialized in popular song has two accepted pronunciations. The number enclosed in parentheses is the above-mentioned index:

tomato(01) T AX M EY T OW

tomato(02) T AX M AA T OW

(These pronunciations are rendered in the “ARPAbet” phonetic alphabet.)

A “decode span” or decode acoustic span” is the same as a “full span” or “full acoustic span”.

A “feature vector” is a multi-dimensional vector, with elements that are typically real numbers, comprising a processed representation of the audio in one frame of speech. A new feature vector may be computed for each 10 ms advance within the source utterance. See “frame.”

A “frame” is the smallest individual element of a waveform that is matched by an ASR system’s acoustic model, and may typically comprise approximately 200 ms of speech. For the purpose of computing feature vectors, successive frames of speech may overlap, with each new frame advancing, e.g., 10 ms within the source utterance.

A “full span” or “full acoustic span” is the entire audio segment decoded by a secondary recognition step, including the audio of acoustic prefix words and acoustic suffix words, plus the putative target span.

A “grammar” is a symbolic representation of all the permitted sequences of words that a particular instance of a grammar-based ASR system can recognize. See “VXML” in this glossary for a discussion of one way to represent such

a grammar. The grammar used by a grammar-based ASR system may be easy to change.

“Grammar-based ASR” is a technology for automatic speech recognition in which only the word sequences allowed by suitably specified grammar can be recognized from a given audio input. Compare with “open dictation ASR.”

The variable “h” refers to a “history” or “language model context,” typically comprising two or more preceding words. This functions as the conditioning information in a language model probability such as $p(w|h)$.

“NLU” refers to natural language understanding: the automatic extraction, from human-readable text, of a symbolic representation of the meaning of the text, sufficient for a completely mechanical device of appropriate design to execute the requested action with no further human guidance.

An “NLU confidence score” is a numerical score that reflects the strength of evidence for a particular NLU meaning hypothesis.

“Open dictation ASR” is a technology for automatic speech recognition in which in principle an arbitrary sequence of words, drawn from a fixed vocabulary but otherwise unconstrained to any particular order or grammatical structure, can be recognized from a given audio input. Compare with “grammar-based ASR.”

A “placeholder” or “placeholder word” is the same as an “aggregate” or an “aggregate word.”

A “phonetic alphabet” is a list of all the individual sound units (“phonemes”) that are found within a given language, with an associated notation for writing sequences of these phonemes to define a pronunciation for a given word.

A “primary recognition step” is a step in the operation of some embodiments, comprising supplying a user’s spoken command or request as input to the primary recognizer, yielding as output one or more transcriptions of this input, labeled with the start time and end time, within this input, of each transcribed word.

A “primary recognizer” is a conventional open dictation automatic speech recognition (ASR) system, in principle capable of transcribing an utterance comprised of an arbitrary sequence of words in the system’s large but nominally fixed vocabulary.

A “proper name” or “proper name entity” is a sequence of one or more words that refer to a specific person, place, business or thing. By the conventions of English language orthography, typically the written form of a proper name entity will include one or more capitalized words, as in for example “Barack Obama,” “Joseph Biden,” “1600 Pennsylvania Avenue,” “John Doe’s Diner,” “The Grand Ole Opry,” “Lincoln Center,” “Café des Artistes,” “AT&T Park,” “Ethan’s school,” “All Along the Watchtower,” “My Favorite Things,” “Jimi Hendrix,” “The Sound of Music” and so on. However, this is not a requirement, and within the context of this specification purely descriptive phrases such as “daycare” or “grandma’s house” may also be regarded as proper name entities.

A “secondary recognition step” is a step in the operation of some embodiments, comprising supplying a selected portion of the user’s spoken command or request as input to the secondary recognizer, yielding as output one or more transcriptions of this input, each transcription possibly labeled with (1) a confidence score and (2) one or more associated meaning variables and their values.

A “secondary recognizer” is an automatic speech recognition (ASR) system, characterized by its ability to perform very rapid adaptation to new vocabulary words, novel word

sequences, or both, including completely novel proper names and words. A secondary recognizer generates an ASR confidence score for its output, and may be operated in “n-best mode” to generate up to a given number n of distinct outputs, each bearing an associated ASR confidence score.

The term “semantics” refers to (1) of or pertaining to meaning, as extracted by the NLU system, (2) the set of possible meanings that may be extracted by the NLU system, taken as a whole.

A “span” is a contiguous section of the input utterance, identified by its start time and end time within the whole of the input utterance (hereafter called the “span extent”), hypothesized to comprise a proper name entity, and labeled with the putative type of this entity (hereafter called the “span type”). The term may also include acoustic prefix and suffix words, not nominally part of the proper name entity per se. See also “acoustic prefix” “acoustic suffix”, “target span” and “full span.”

A “span extent” is the start time and end time of a span, within an input utterance.

A “span type” is the putative type of the proper name entity believed to be present within the span; thus a personal name, business name, numbered street address, etc.

A “target span” is the portion of the acoustic span, decoded by a secondary recognition step, that nominally contains the words of the proper name entity. Thus, the term refers to the acoustic span, exclusive of the acoustic prefix words and acoustic suffix words.

An “understanding step” is a step in the operation of some embodiments, comprising supplying as input the text and word timings of the user’s utterance as generated by the primary recognizer, and yielding as output one or more hypothesized symbolic meanings of the user’s input, each such meaning possibly including the identification of one or more acoustic spans, comprising a span extent and span type, each such span to be separately processed by a secondary recognition step. Optionally, each hypothesized symbolic meaning may include an associated NLU confidence score.

An “utterance” is audio presented as input to an ASR system, to be transcribed (converted into text) by that system.

A “vocabulary” is, informally, a list of the words with associated pronunciations, which forms part of the input to an ASR system, and which defines the words that could in principle be recognized by such a system. Formally, the term may refer to a list of baseforms.

“VXML” is a popular standard for specifying the grammar, for grammar-based ASR systems.

The variable “w” refers to a generic word, including an aggregate word. The textual output marking a period of silence, in a transcription generated by an ASR system, is also regarded as a word.

DETAILED DESCRIPTION

Embodiments Overview

Various examples of the disclosed techniques will now be described in further detail. The following description provides specific details for a thorough understanding and enabling description of these examples. One skilled in the relevant art will understand, however, that the techniques discussed herein may be practiced without many of these details. Likewise, one skilled in the relevant art will also understand that the techniques can include many other obvious features not described in detail herein. Additionally,

some well-known structures or functions may not be shown or described in detail below, so as to avoid unnecessarily obscuring the relevant description.

The terminology used below is to be interpreted in its broadest reasonable manner, even though it is being used in conjunction with a detailed description of certain specific examples of the embodiments. Indeed, certain terms may even be emphasized below; however, any terminology intended to be interpreted in any restricted manner will be overtly and specifically defined as such in this section. One will recognize that the same thing may be explained in many different ways and that different portions of the specification may elaborate upon the same or similar concepts.

Various of the disclosed embodiments attain high recognition accuracy and understanding of freely spoken utterances containing proper names such as, e.g., names of persons, streets, cities, businesses, landmarks, songs, videos or other entities that are known to be pertinent to a particular user of such a system. Various embodiments augment the recognition system with methods that recognize and understand completely novel proper names, never before incorporated into the system in question. Various embodiments may achieve this benefit with extremely low latency, e.g., on the order of a few hundred milliseconds.

Some embodiments may be used to recognize entities, such as numbered street addresses, or street intersections, that include within them street names and possibly city and state names as well. An example of the latter would be “333 Ravenswood Avenue” or the more precise “333 Ravenswood Avenue, Menlo Park, Calif.” Word sequences that are purely descriptive and generic, such as “grandma’s house,” “the office,” “daycare,” “the playground” and so on, which the user has identified to the system as personally significant, may also be addressed in some embodiments. Throughout this document, the terms “proper names” and “proper name entities” will be understood to refer to the proper names and word sequences discussed in this and the preceding paragraph.

Some embodiments also extract a symbolic meaning, as appropriate, associated with the identity or relevant particulars of the recognized entity (such as the index of a particular entry in a list of businesses or personal contact names, the number portion of a street address, the user’s current address or the internal symbolic label of a street name within an automatic mapping or navigation system), so that the system as a whole may respond appropriately to the user’s spoken request.

An additional benefit of some embodiments is higher accuracy recognition of proper name entities than can be achieved with conventional methods, such as direct adaptation of an open dictation ASR system. This benefit may be obtained because some embodiments place additional sources of information at the disposal of the speech decoding and meaning assignment process. This information may be principally but not exclusively derived from an intermediate NLU processing step, from the state of the system as a whole, such as recent prior user inputs and search results, or information about or associated with the user, such as the contents of a personal or professional calendar.

An additional benefit of some embodiments is that the open dictation ASR component of the system, prepared by some of the methods described here, may require no further adaptation or modification to enable recognition of names and entities that are not initially present in its vocabulary. Thus, this open dictation ASR component may be shared by a multitude of users, with the necessary adaptations to

enable recognition of proper names confined to other components of the system. This may provide several important advantages.

First, adaptation of open dictation ASR systems is generally a time-consuming process, executed over the course of several hours if not days or weeks. Such computational requirements may make it infeasible to rapidly modify an open dictation ASR system to enable recognition of, for instance, personal names that a user has just entered into a contact list, business names in an event or appointment record that a user has just entered into an electronic calendar, or street names of a region that a user is currently navigating or intends to navigate. By contrast, various embodiments allow the system as a whole to be adapted effectively instantaneously to such names.

Second, such adaptation of open dictation ASR systems typically involves preparation of a new vocabulary, language model and acoustic model, or some subset thereof, each of which is an electronic computer file. Such files can be large even by current standards of electronic storage technology. For instance a typical language model may occupy some 4 GB of storage. The computational and hence economic cost to prepare these specially adapted files, and the associated economic cost to store them and load them on demand, for each individual user of the system, may be prohibitively high. By contrast, some embodiments do not require adaptation of the open dictation ASR system, yet yield accuracy akin to or superior to the performance of an open dictation ASR system that has been adapted in the conventional fashion.

Third, in some embodiments a “primary” open dictation ASR system resides at a central server, whereas a “secondary” grammar-based ASR system resides within a smart-phone, automotive dashboard, television, laptop or other electronic computing device that is the user’s personal property. The latter device is referred to as the “client device” or the “client” herein. The system adaptations to enable recognition and understanding of the proper name entities associated with that particular user may be confined to the secondary ASR system and may be executed exclusively within the client device in some embodiments.

This means that in effecting this adaptation, the user’s privacy is not compromised by the transmission, storage and processing of the contents of the user’s personal address book at a central server. As every new month brings news of breaches of commercial and governmental “high security” computer systems, an architecture that ensures that the user’s personal information is never sent to other devices is highly desired.

System Overview

Various embodiments accept as input an audio signal comprising fluent, natural human speech, which notably may contain one or several proper names, or unorthodox sequences of otherwise ordinary words. The embodiments may produce as output an accurate textual transcription of this audio signal, and optionally a symbolic rendering of its meaning.

In some embodiments, the system comprises four major functional components, respectively a primary speech recognizer (or more simply a primary recognizer), a natural language understanding module (also called an NLU module, language understanding module or just understanding module), an adaptation object generator, and a secondary speech recognizer (or secondary recognizer). Some embodiments also include a fifth major functional component, the score fusion and hypothesis selection module, which will be discussed in later sections. The system as a whole may

include mechanisms to cause these components to operate and communicate as described herein, and to store the input audio signal in such form that it may be reprocessed, in whole or part, during the operation of some embodiments.

The primary recognizer may include a conventional open dictation automatic speech recognition (ASR) system. Such a system accepts as input an audio signal comprising human speech. It may produce as output a textual transcription of this input, labeled with the start time and end time, within the input audio signal, of each transcribed word. It may also attach an ASR confidence score to each transcribed word and optionally to the output transcription as a whole. The primary recognizer may be an “open dictation” ASR system in that it may transcribe an utterance comprising an arbitrary sequence of words that belong to its vocabulary. This is contrasted with a grammar-based ASR system that can recognize only certain predetermined word sequences. In those embodiments where the recognizer is “conventional”, this designation is used in the sense that the recognizer does not make use the embodiments described herein. As a result, the primary recognizer may be assumed to have a large but fixed vocabulary.

This vocabulary may be difficult or impossible to augment with proper names or other novel words not presently in the vocabulary. Attempting to do so may require many minutes, hours or possibly even days of computational effort. These unknown proper names or other novel words—“unknown” in the sense of not being listed in the fixed vocabulary—are therefore not recognizable by this primary recognizer. Moreover, if presented with an audio signal comprising words that belong to the vocabulary, but which are spoken in an unusual and possibly nominally meaningless sequence, such as “The The” (the name of an English musical group founded in 1979), the primary recognizer may have difficulty generating a correct transcription. Again, it is often difficult for the primary recognizer to accurately transcribe such unorthodox word sequences without significant computational effort.

In some embodiments, the natural language understanding module accepts as input the transcription and word timings generated by the primary recognizer and emits as output one or more hypotheses of the meaning of the utterance (also called an NLU hypothesis, meaning hypothesis or just a meaning). This meaning may be represented in a symbolic form suitable for processing or execution by a computer. Each meaning hypothesis optionally includes a numerical NLU confidence score, which reflects the strength of evidence for that particular meaning.

This module may identify a particular word or word sequence in the input transcription that potentially comprises a proper name entity and label this word or word sequence with a putative type (for instance, a person’s name, a street intersection, a numbered street address, and so on). Each such word or word sequence is called a proper name entity acoustic span, or the acoustic span of a proper name entity, or just an acoustic span. The basis for marking this word or word sequence as an acoustic span may be quite indirect, and may not reflect the nominal meaning of the words that comprise it.

A given hypothesis may include one or more such acoustic spans, each one constituting an information element that must be resolved to fully specify the meaning of the phrase. The transcription and meaning of the span may be determined by the context in which the embodiment is applied. In some embodiments, a given hypothesis may not include any acoustic spans at all. In this case the proper name recognition embodiments discussed herein may not apply.

The adaptation object generation module or adaptation object generator creates computational objects that are used to adapt the secondary recognizer in the manner described in the next paragraph. As detailed herein, this process may be divided into two stages, respectively object preparation and object finalization.

Finally, the secondary recognizer also comprises an ASR system, insofar as it accepts an audio signal as input and generates a transcription, and other information, as output. It may also attach an ASR confidence score to each transcribed word and optionally to the output transcription as a whole; it may also be operated in “n-best mode,” to generate up to a given number *n* of distinct outputs, each bearing an associated ASR confidence score. However, its characteristics may be markedly different from those of the primary recognizer. Specifically, the secondary recognizer may be capable of very rapid adaptation to new vocabulary words, novel word sequences, or both, including completely novel proper names and words. Here “very rapid” may be “performed typically in a few hundred milliseconds or less.” This adaptation is typically achieved by loading the secondary recognizer with an adaptation object generated by the adaptation module. The secondary recognizer may also be unlike the primary recognizer as it is constrained to transcribe only a relatively small collection of phrases, numbering, e.g., in the tens, hundreds or thousands, rather than, e.g., the billions of phrases supported by the primary recognizer.

The primary and secondary recognizers may further be distinguished based upon their usage. The secondary recognizer, rather than processing the audio signal comprising the entirety of the user’s spoken input, may operate upon only one or two short segments of the signal, extracted, e.g., from a saved copy of the signal. These segments are referred to herein as acoustic spans (or simply spans).

Acoustic Spans

The primary recognizer described above may resemble a large-vocabulary open dictation ASR system, which is founded upon the Bayesian minimum-error decoding equation $W^* = \operatorname{argmax}_W P(A|W)P(W)$, wherein *A* is the audio signal to be decoded (transcribed), *W* is an hypothesis (guess) as to the correct decoding (transcription), *W*^{*} is the final decoding (transcription), and *P*(*A*|*W*) and *P*(*W*) are the numerical values of the acoustic model and language model respectively, for the indicated inputs. Such a system may derive its generality from the use of a statistical language model to compute the quantity *P*(*W*). But this very generality can imply the intractability to modification or adaptation that is one of the primary recognizer’s characteristics. Likewise the features and performance of the adaptation object generator may be those of a grammar compiler; similarly the features and performance of the secondary recognizer may be those of a grammar-based ASR system.

It should be noted that these are only examples of the ASR technology found in some embodiments and the embodiments should not be understood to be restricted to such instances. It is the capabilities of the primary and secondary recognizers that are important, not the specific means by which their functions are achieved. However, in some embodiments the primary recognizer is a large-vocabulary open dictation ASR system that uses statistical language models as described above and the secondary recognizer is a grammar-based ASR system. Many of the examples presented herein will proceed on this basis.

Example Standard Processing Overview

The following is an example application of some embodiments. This example considers the processing of two commands, the first of which does not employ various of the

disclosed embodiments, and the second which does. One will recognize this discussion merely as exemplary so that the reader may compare and contrast the two sequences of operations. Consider a system with a spoken language interface, running on a portable, Internet-connected tablet computer, that enables a user to search for businesses by category or service, and if desired, to inquire for additional pertinent information about individual enterprises that have been retrieved.

Suppose for example that the user is located in the Golden Gate Heights neighborhood of San Francisco and issues the spoken command “find some nearby Indian restaurants.” The following description explains how this initial command, which contains no proper name entities, is transcribed and analyzed for meaning. That is, as this command includes no proper names or proper name entities, it need not employ various of the disclosed embodiments. Accordingly, the following operations may be performed by the primary recognizer and natural language understanding modules respectively without using the disclosed embodiments.

First the audio input is supplied to the primary recognizer, which emits a transcription comprising a sequence of words in the primary recognizer’s vocabulary. It is assumed that all the words in the sample command above are in this vocabulary. Next this transcription is supplied as input to the natural language understanding module. Suppose that no proper name entity acoustic spans are identified and that only a single meaning hypothesis is generated. This yields a symbolic representation of the command’s meaning

```
YelpSearch [m_latitude=37.755196, m_longitude=
122.46853, m_sortBy=PROXIMITY, m_categories=
[indpak], m_maxResults=5]
```

which the system then converts into a query of the “Yelp®” online business search service. This query is passed to the Yelp® service, which returns a list of five restaurants that serve Indian cuisine, scattered around the indicated latitude and longitude, which correspond to the user’s current location. The system then uses other online services, and its inbuilt logic, to create the final response to the user. The contents of the tablet display, listing the five retrieved restaurants, and showing their locations on a map, appears in FIG. 1.

Referring to FIG. 1, the businesses retrieved are “Masala Dosa,” “Tikka Masala,” “Guddu de Karahi,” “Naan-N-Curry,” and “Noori Pakistani & Indian Cuisine.” Some of the words in these business names, for example “cuisine,” “curry,” “Indian” and “Pakistani,” would be widely agreed to belong to standard American English, and hence are likely to be found in the vocabulary of any reasonably robust ASR system. However others, such as “masala” and “tikka,” are more doubtful. “Karahi,” though in fact nominally an English language word, insofar as it is present in some dictionaries, is so unusual that it is unlikely to be found in any ASR system vocabulary. And “Guddu,” which is a given name for a male child on the Indian subcontinent, is almost guaranteed not to be found in any commercial ASR system for the North American market.

Although the system may not include any of “de,” “Dosa,” “Guddu,” “Karahi,” “Masala,” “Noori” or “Tikka” in its ASR vocabulary, the system implementing the present embodiments is nevertheless capable of recognizing and responding properly to a command like “tell me how to get to Guddu de Karahi.” This may be achieved in some embodiments by creating a specialized recognizer that can process the indicated business names (and in some embodiments nothing else), exploiting the information obtained by both the primary recognizer and the natural language mod-

ule, and other such information that may be relevant, so that an appropriate acoustic span may be identified, and then deploying this specialized recognizer to good effect.

First, anticipating that the succeeding command may make reference by name to one of the businesses listed on the tablet display, the system operates the adaptation object generator to create an object suitable for adapting the secondary recognizer to recognize precisely these names. This may be done by preparing a grammar, illustrated in graphical form in FIG. 2, that contains exactly these names, and compiling it into a binary form so that it is ready for use by the secondary recognizer. This operation, which may typically take a few hundred milliseconds, may be performed immediately upon receiving from Yelp® the list of names to be shown on the tablet display. Compilation may involve (1) obtaining one or more pronunciations for each indicated word in the grammar (this may typically be done by first searching a vocabulary, but if this search fails any required pronunciations may be automatically generated by a “grapheme to phoneme” or “g2p” processing module, which applies the standard rules of English language pronunciation to the given word spelling to produce one or more plausible pronunciations), (2) creating a computational structure that permits words to be decoded only in the order allowed by the grammar, (3) attaching to this structure operations to be performed on indicated meaning variables when a given decoding is obtained (which may typically comprise assigning values to these variables), and (4) emitting this structure in such form that it may be immediately loaded by a suitable grammar-based ASR system and used to guide its decoding of audio input. This compiled grammar, denoted “business-names.g” in FIG. 3, may be labeled with its type (in this case effectively business-names) and held for possible future use. In some embodiments, this comprises the adaptation object generation step.

It should be noted that this grammar may be created speculatively and that this action may not require any great prescience on the part of the system. In some embodiments, other grammars, for example covering the user’s personal contacts, the businesses in the user’s personal calendar, the artists, song titles, and album names stored in the user’s iPod® or USB flash drive, or all the numbered street addresses for the city in which the user is currently located, may have been created on an equally speculative basis. The net result is that effectively a panoply of specially adapted secondary recognizers may be available for use in the secondary recognition step, to process various spans that may be identified, of various types.

The system may now wait for further input. There is no guarantee that the next command will make reference to one of the businesses. Indeed the user could potentially issue a request for a different cuisine, saying perhaps “actually I feel more like Italian food instead.” As this sentence, like the request for Italian restaurants, lacks any proper name entities, this command would be processed in the conventional manner described above, yielding the symbolic meaning `YelpSearch [m_latitude=37.755196, m_longitude=-122.46853, m_sortBy=PROXIMITY, m_categories=[italian], m_maxResults=5]` which would in turn trigger a similar query for Italian restaurants and on-screen display of those results. Similarly, an entirely different command could be issued, such as “show me some nearby hardware stores,” or “where can I get my shoes repaired,” which would be processed in a like manner.

But for the sake of illustration let us suppose that the user now speaks the command “tell me how to get to Guddu de

Karahi.” As described above, the audio input is first passed to the primary recognizer, which generates an initial nominal transcription of its input, labeled with word timings. This is referred to herein as the primary recognition step. The user’s audio input may also be retained for later processing by the secondary recognizer.

This initial transcription may well be incorrect. Indeed, if the vocabulary of the primary recognizer does not include all the words spoken by the user it is very likely to be incorrect. For the sample command presently under discussion, and one particular audio input comprising a vocalization of this command, the primary recognizer actually produced as output the word sequence “tell me how to get to go to do a call Rocky,” labeled with word timings as shown in FIG. 3.

Next this imperfect initial transcription may be presented to the natural language understanding module. This module processes the input word sequence, and determines by application of standard methods of computational linguistics to the first six words of the transcription—“tell me how to get to”—that the user is making a request for directions. Noting that the rest of the transcription—“go to do a call Rocky”—is both nominally somewhat nonsensical, and also occupies a position in the phrase as a whole that in conventional conversational English would likely comprise the name of the target to be navigated to, the language understanding module also determines that the portion of the audio input corresponding to this part of the transcription probably contains a spoken rendering of one of the displayed business names. As discussed herein this selected portion of the audio input is referred to as the proper name entity acoustic span, or acoustic span of a proper name entity, or just acoustic span for short, that is now to be processed by the secondary recognizer. Note that by virtue of the word timings generated by the primary recognizer, the acoustic span in question is known to begin at 1330 ms into the audio input, corresponding to the start of the word “go,” and end at 2900 ms into the audio input, corresponding to the end of the word “Rocky.” In this way the extent associated to the span has been determined. This entire operation comprises the language understanding step of this example.

The system may then proceed to the secondary recognition step. By virtue of the language understanding module having determined that a particular segment of the audio input probably comprises one of the displayed business names, the already-compiled grammar which enables recognition of these names (and in some embodiments, only these names) is loaded into the secondary recognizer. The acoustic span of the putative business name—that is, the sequence of samples from 1330 ms to 2900 ms of the audio input signal—is presented to the secondary recognizer as its input. As the audio that comprises the extraneous words “tell me how to get to” has been suppressed from the secondary recognizer’s input, and as the secondary recognizer is constrained by its grammar to recognize only the phrases “Masala Dosa,” “Tikka Masala,” “Guddu de Karahi,” “Naan-N-Curry,” and “Noori Pakistani & Indian Cuisine,” the correct transcription “Guddu de Karahi” of the acoustic span is easily obtained. Moreover, associated with the successful recognition of “Guddu de Karahi,” the meaning variable `selected_business_index` is set to the value 3. This completes the secondary recognition step in this example.

With these secondary recognition results in hand, the acoustic span transcription “Guddu de Karahi” may be interpolated into the primary recognizer’s transcription, replacing the word sequence “go to do a call Rocky” that was initially guessed for this span, thereby yielding a final transcription “tell me how to get to Guddu de Karahi.”

Correspondingly, the symbolic meaning directionApnrCommand is populated with a parameter identifying the navigation target, yielding the complete symbolic meaning directionApnrCommand [selected_business_index=3].

This symbolic meaning may then be processed by other functional elements of the system, extracting information (including location) pertinent to the third business from the five-element array of such objects, executing appropriate operations to find a route from the user's current position to the indicated location, rendering a map showing this route, etc. The map may depict other associated information as deemed useful and pertinent to the application context by the system designers. The resulting image in this example is shown in FIG. 4.

As this example demonstrates, the disclosed embodiments provide many advantages over conventional systems. The "Achilles heel" of grammar-based ASR technology is that the user must speak within the grammar or the technology will not function. In contrast, the disclosed embodiments do not comprise simply causing the user to stay within a grammar, when speaking his or her request.

The disclosed embodiments allow the user to speak freely, using the words and phrase structure that come naturally when expressing the desired action. The associated audio input may then be analyzed by the primary recognizer and the language understanding module to determine if a proper name entity, substantive to the correct processing of the command, has in fact been spoken. If so the proper name entity's extent within the audio input, and putative type, are identified, and this specific segment of audio may then be processed by the secondary recognizer, adapted to recognize the proper name entity within a relatively small list of possibilities. This narrowing of the task, in two important senses—first by pruning away the freely formed and now-extraneous audio that would confound a grammar-based ASR system, and second by adapting the secondary recognizer to drastically reduce the space of possible transcriptions may allow the secondary recognition step to succeed. This analysis and subsequent narrowing may in turn depend upon the ability, afforded by various disclosed embodiments, to integrate information and insights normally outside the scope of ASR technology—specifically in this case that a prior command generated a list of businesses and hence that a followup command naming one of them is not unlikely, plus the observation that the phrase "tell me how to get to," or a myriad of other phrases of similar meaning, was probably followed by a proper name or other vocalization of a navigation target.

One will recognize that many methods exist by which the primary recognizer and language understanding module determine that an acoustic span appropriate for processing by the secondary recognizer is present (some examples are provided below). Many of these different methods may be used. Any suitable information and methods known in the art may be employed if they are effective to identify spans, adapt the secondary recognizer, and to then operate on each span with an appropriately adapted secondary recognizer. Indeed, by repurposing existing tools to perform the above functions the embodiments may be readily implemented in existing designs.

Some embodiments may even include several distinct and competing mechanisms for identifying acoustic spans, with all of them processed by distinct and separately adapted secondary recognizers, with a final determination of one or a few surviving hypotheses (surviving for presentation to

and ultimate disambiguation by the user) performed by the "score fusion and hypotheses selection" module discussed herein.

Detailed Discussion of Features

The following sections present a more precise description of various embodiments, explain important ways in which they may be applied to good effect, describe particular architectures that are matched to specific applications, detail techniques for overcoming various obstacles, explain how to deal with errors in the identification of the extent or type of a span, and describe methods for selecting between multiple competing transcriptions and meanings of a user's spoken input.

Various embodiments divide the speech decoding process (and as we shall see, the meaning extraction or "understanding" process as well) into a primary recognition step, one or more understanding steps, one or more adaptation object generation steps (which may comprise two stages, an object preparation stage and an object finalization stage), one or more secondary recognition steps, and (optionally) a score fusion and hypothesis selection step.

The primary recognition step comprises recognition of the input utterance by a conventional open dictation ASR system, though this system may have been specially prepared to assist the language understanding module to identify the extent and type of one or more acoustic spans. This yields a transcription of the utterance (and possibly alternate transcriptions as well) into the vocabulary of the open dictation recognizer, plus nominal start and end times for each transcribed word.

The primary recognition step, which may be accomplished by the nominally more powerful, more flexible and more computationally demanding primary recognizer (the object of the comparative "more" being the secondary recognizer), may not bear the full responsibility of generating the final transcription of the input utterance. Instead, the main objective of this step may be to provide a sufficiently accurate transcription for the language understanding module to do its work of hypothesizing one or more symbolic meanings for the user's command, including the extent and type of any proper name entity acoustic spans that may figure in the full specification of this meaning. It should be clear from the example in the overview that the words in the primary recognizer's transcription may be far from correct. In fact the primary recognizer may provide several alternate transcriptions of the input waveform, each one subject to the processing steps described below; a means for selecting the final preferred transcription and its associated meaning will be explained shortly.

The output of the primary recognition step, comprising (1) a nominal transcription, (2) the start and end time within the waveform of each transcribed word at the granularity of a single frame and (3) possibly information, described further below, of use in determining the extent and type of any acoustic spans, may then be passed to the understanding step.

The understanding step applies the methods of natural language understanding to hypothesize one or more symbolic meanings for the nominal transcription and as appropriate to identify the extent and type of any proper name entity acoustic spans that contribute to this meaning. Each acoustic span becomes an element of the hypothesis, to be processed by an associated secondary recognition step to yield the span's transcription and meaning.

At this point relevant, but heretofore unexploitable information, may now be incorporated into the speech recognition and meaning extraction process. Notably, information

that is derived from any portion of the primary recognition transcription of the current utterance, from previously decoded utterances, or even from wholly non-linguistic information sources, such as the physical location of the user as determined by GPS, or the known preferences or characteristics of the user, can be exploited at this stage to assign span extents and types, and appropriately adapt the secondary recognizer. The adaptation of the secondary recognizer, or rather the preparation of one or more objects that may be used to adapt one or more instances of the secondary recognizer, comprises the adaptation object generation step. As noted in the example above, various embodiments may speculatively create or make use of various adaptation objects, appropriate to the type or types of spans to be processed. This adaptation may comprise preparing the secondary recognizer to recognizer completely novel words, restricting the secondary recognizer so that it does not use certain other words in its vocabulary or uses them only in particular orders, or both.

With regard to the output of the understanding step, the output may comprise a collection of hypotheses, each one containing one or more acoustic spans. We now pass to the third step, which is the secondary recognition step. In fact there may be many such steps, as each such span is now decoded by a grammar-based speech recognizer, which has been specially adapted to the span type. For whereas the adaptation of an open dictation recognizer to a specialized vocabulary or context is typically computationally expensive, a new grammar can be generated or compiled, or a grammar with unpopulated placeholder "slots" can be completed and made ready for service in a few hundred milliseconds or less in some embodiments. The output of this secondary recognition step, performed solely on the subject acoustic span, using a suitably specialized grammar, may be taken as the nominal transcription of the span. We note at this point that the literal sequences of this same grammar may be labeled, in appropriate and conventional ways, with the meaning of each potential decoding path through the grammar. Thus the act of transcribing the span may at the same time generate an appropriate symbolic meaning, associated to the transcription.

Thus, while grammar based ASR fails when presented with freely-formed human speech, which typically lies outside the scope of even elaborate grammars, the disclosed embodiments perform well when presented with in-grammar utterances. The preceding processing stages establish this desired condition. Consider an acoustic span which is known, or more correctly hypothesized by earlier processing steps, to consist of the name of one of a few businesses drawn from those listed in a user's daily appointment calendar for a particular day. If that sole portion of the original utterance is provided as the audio input to a grammar-based ASR system, and the grammar used for decoding comprises all and only the business names extracted from the user's calendar for that day, then it is highly likely that the correct proper name will be decoded.

Such secondary recognitions may be performed for each of the acoustic spans identified by the prior decoding stages, until a final transcription is obtained for the whole of the original utterance. If no competing alternative meaning hypotheses were proposed by the prior processing steps, then the decoding is complete. However, this may not always be the case. More likely, several alternative transcriptions, each with one or more associated meaning hypotheses, may have been generated, each hypothesis having NLU and ASR confidence scores. It remains to select the final preferred decoding, or at a minimum, assign a

confidence score to each whole decoding, and provide a ranked list of alternatives. As differing hypotheses may comprise different numbers of acoustic spans, this may force the comparison of hypotheses that are based upon different numbers of confidence scores. One will recognize various approaches to combine such scores in a consistent manner, to allow meaningful and reliable score-based ranking. The NLU system itself may be involved in generating this ranking.

This is the function of the final step, of score fusion and hypothesis selection. In this step, a special "complete hypothesis ranking grammar" is prepared, in which each admissible path comprises the entirety of one individual complete transcription. The original utterance in its entirety may then be decoded against this grammar by the secondary recognizer operating in n-best mode, yielding an acoustic confidence score for each complete hypothesis, nominally expressed as $P(T_i|A)$. Here T_i is the text associated with the i th hypothesis, and A is the acoustic input, which is constant across the hypotheses being ranked. It is possible that this will suffice, and a ranking of hypotheses may be made purely upon this acoustic score.

However, if the NLU confidence scores can be normalized to probabilities, they may be meaningfully combined with the ASR confidence scores by the following application of the laws of conditional probability. Let T_i and A denote the transcription and acoustic input as above, and let M_i denote the symbolic meaning assigned by NLU processing to the i th hypothesis. Write $P(M_i|T_i)$ for the NLU confidence score of the i th hypothesis meaning, given the associated transcription. Then by the product law for conditional probabilities we have $P(M_i, T_i|A) = P(M_i|T_i, A)P(T_i|A)$. Now make the reasonable approximation $P(M_i|T_i, A) \approx P(M_i|T_i)$ which expresses mathematically the concept that the meaning of the user's spoken input is accurately conveyed by a correct transcription of the user's speech. This then yields $P(M_i, T_i|A) \approx P(M_i|T_i) P(T_i|A)$, which expresses our confidence that we have the correct meaning and transcription of the (fixed) acoustic input A .

Example Applications

We now discuss various applications of the disclosed embodiments, and show how it may be used to achieve certain desired functions. For example, the disclosed embodiments may apply, e.g., to: business names (resulting from a search); business names (retrieved from a personal phone book, personal calendar, or both); personal contact names (retrieved from a phone book, or from a calendar); locations (numbered street addresses); locations (intersections); locations (landmarks); music library search; and video library search.

In the first case of business names the adaptation object may be constructed from the names retrieved in the just-executed search. In the second case of business names the adaptation object may be constructed from business names retrieved from a personal phone book, personal calendar or both, possibly restricted to the current day's personal calendar. In the personal contact names case the adaptation object may be constructed from personal contact names retrieved from a personal phone book, personal calendar or both, possibly restricted to today's personal calendar. In the first locations case the adaptation object may be one of many constructed well in advance, each object comprising the valid street addresses for each street in every political subdivision in a country (typically a city), the adaptation object actually used being determined either by the user's current location as determined say by GPS, by an explicit or implicit preceding request for a particular such subdivision,

or by the identity of the political subdivision as decoded by the primary recognizer from some part of the user's utterance (viz the transcribed words "Menlo Park" in the primary recognizer transcription "tell me how to get to three thirty three Ravenswood Avenue in Menlo Park").

In the second locations case the adaptation object may likewise one of many constructed well in advance, each object comprising intersections of each street in every given political subdivision in a country, with the adaptation object actually used determined as described above. In the music library search case various adaptation objects are constructed from the artist names, song names, album names, and genre names in a user's personal music storage device. In the video library search case various adaptation objects are constructed from the actor names, director names, and genre names in a given catalog of video content to be navigated. However this list is intended to be merely exemplary and not exhaustive. The disclosed embodiments may be applied in other ways as well.

Incorporating Acoustic Context

As is well known to phonological linguists and developers of ASR technology alike, the words spoken before and after any given word in a fluently spoken phrase may have a significant impact on the pronunciation of the given word. This is the phenomenon of coarticulation. Two varieties are commonly recognized: anticipatory coarticulation, wherein the word or words spoken before a given word change its pronunciation, and carryover or perseverative coarticulation, wherein the word or words spoken after a given word likewise change its pronunciation.

The causes of coarticulation are complex, but the effect is due at least in part to the fact that the elements of the human speech apparatus (lips, jaw, tongue, vocal folds, etc.) are physical objects with non-zero mass, which therefore cannot be moved instantaneously from one configuration to the next. Thus the position of this apparatus before the speaker produces a given phoneme, and likewise the intended position that is coining next, will affect the position used when producing the phoneme and hence the sound itself.

Allophones may be used to address this phenomenon, wherein the templates or models used to match a particular phoneme are made to depend upon the sequence of phonemes that precede it, and those that follow it. However the secondary recognition step, or more properly the generation of the adaptation object associated to this step in some embodiments, does not account for coarticulation. In the example given above, the adaptation object—in the example, a grammar of business names—that was prepared shows no words either preceding or following the listed names. Such a grammar would be appropriate for decoding speech that consists of one of these names having been spoken in isolation, with no preceding or following words. But in fact the speech to be decoded is an extract from a longer, fluently spoken phrase. Though in the example only silence followed the spoken name, in fact several words precede it, and these are likely to influence the speaker's pronunciation of one or more of the words within the business name itself. If this acoustic context is not somehow reflected in the recognition process, it is possible that the secondary recognition will fail.

Accordingly, some embodiments employ a method for incorporating acoustic context into the adaptation object generation step, so that the secondary recognizer may accommodate coarticulation effects. The following discussion is a running example that demonstrates the operation of one embodiment of the method. The example is discussed in relation to the FIGS. 5, 6, and 7.

Let us suppose that the user speaks the command to the system "send a message to Steve Youngest hi Steve how are you." The primary recognizer transcribes this audio input, producing the nominal output "send a message to steve young us hi steve how are you." The natural language module receives this transcription and associated word timings, and on the basis of the evidence presented decides that the words "steve young us" comprise an acoustic span of type personal-contact-name, with extent from 1690 ms to 2530 ms within the input utterance.

The nominal adaptation object, comprising the contact names say in the user's address book, has the structure illustrated in FIG. 5. That is to say, it comprises a list of alternatives, each one a personal contact name, and each labeled with some suitable meaning variable command to be executed if the associated literal sequence is decoded by the secondary recognizer. But it contains no information about the acoustic context in which the contact name was spoken. Absent such information, when the grammar is compiled, it is impossible to adjust the phoneme models used to represent each contact name literal to account for the fact that the spoken name to be decoded was preceded by the fluently spoken words "message to," and followed by silence and then the spoken word "hi" (Note that the fact that silence follows the spoken name may itself be informative to the grammar compilation process, as it suggests the trailing phonemes of the verbalized contact name were not in fact subject to anticipatory coarticulation).

However this precise information can be exposed to the compilation process by expanding the span extent to include some number of the words preceding the putative contact name, and likewise the words that follow it. This is illustrated in FIG. 6, which also introduces some new nomenclature. Here the embodiment has chosen to incorporate into the grammar the two words before the name, "message to," and the two words after then name, "SIL hi" (note in passing that a silence is itself considered a word, here denoted "SIL" in the transcription by the primary recognizer). As illustrated in FIG. 6, these words are called respectively the acoustic prefix and acoustic suffix (equivalently the "prefix acoustic span," "prefix span" or "precontext"; and likewise the "suffix acoustic span," "suffix span" or "postcontext"). The putative contact name itself—what was previously just the "span"—is now denoted the "target acoustic span," the "target span" or just the "target" for short. Finally, if we wish to make clear that we are alluding to the now expanded extent of the span in full, including all three of the prefix, target and suffix, we will refer to the "full acoustic span" or "decode acoustic span" (or just "full span" or "decode span" for short).

Returning now to the compilation process, the adaptation grammar is revised to the structure shown in FIG. 6. That is, the primary recognizer's decoding of the acoustic prefix are inserted as literals along arcs that lead into the previous structure, and the words decoded for the suffix are likewise inserted along arcs that lead out of it. In this example, these literals do not bear any meaning variables, because they serve no function in determining the meaning of the span. Rather, their function is to provide the necessary acoustic context to the compilation process, in selecting the allophones to associate to the pronunciations of each of the personal contact names within the target section of the grammar.

In this embodiment example, when the associated secondary recognition step is performed, the audio excerpt that is processed comprises the extent of the full span, from the nominal start of the first word of the acoustic prefix through

the nominal end of the last word of the acoustic suffix (this is why this is also referred to herein as the “decode span,” because the full extent is in fact processed—decoded—by the secondary recognizer). As there are no alternative pathways in the grammar for the prefix and suffix, this has the effect of causing the secondary recognizer to perform a forced alignment between the prefix literals and their corresponding audio, and likewise between the suffix literals and their audio.

This example procedure has another important benefit as well. It can sometimes be difficult, for either a computer or a trained human, to decide exactly where within a waveform one word ends and the next word begins. Consulting FIG. 6, does the first instance of the word “steve” begin exactly at 1690 ms? Or does it start perhaps one or two frames earlier, at 1680 ms or 1670 ms? Or possibly a bit later, at 1700 ms or 1710 ms? Similar questions arise for the end of the word “us,” which defines the end of the target span. If the secondary recognizer processes only the target span extent, then the associated word timings may need to be highly accurate, perhaps unreasonably so.

By comparison, including the prefix and suffix literals in the grammar, and enlarging the span extent to include the corresponding audio, displaces the problem of finding the span boundaries to the start of the prefix and end of the suffix respectively. Minor errors in assigning these boundaries may be of no consequence, because no substantive decoding, in the sense of choosing among alternatives, is being performed. If a few frames of the start of the word “message” are missing from the span, or those of the end of “a” erroneously included, the forced alignment of the acoustic prefix literals will nevertheless succeed. The same may be true of the suffix. But, in this example, all the frames that comprise the target are more assuredly available and the recognizer can freely compute the best acoustic match between the alternative literals in the target section of the grammar and the associated audio signal.

In this embodiment example, the adaptation object—that is, in this embodiment, the personal contact name grammar—may be populated with the prefix and suffix words as determined by the primary recognition step. This would seem to present a challenge to the desire to achieve low latency decoding of the user’s spoken phrase, as part of the adaptation step is now executed between the language understanding step and one or more of the secondary recognition steps.

Various embodiments contemplate a grammar with so-called “slots,” which are placeholders for literals to be populated at the very last moment, with very low latency. This “slotted grammar,” with a target section comprising the names of the user’s personal contact list, and with four unpopulated slots for the acoustic prefix literals and acoustic suffix literals, may be speculatively created at the system’s leisure as soon as this name list is available. This grammar is illustrated in FIG. 7. Creating and compiling this grammar, leaving the slots unpopulated, is the preparation stage of the adaptation object generation step. It is then held ready for use at the appropriate moment. This slotted grammar may then be populated with the appropriate words extracted from the primary recognizer’s transcription, in the finalization stage of the adaptation object generation step.

Example Client and Server Topology

FIG. 8 is a block diagram depicting various components in a speech processing system **800** having server and client proper name resolution modules as may occur in some embodiments. The depicted topology is merely an example provided for purposes of explanation and one will recognize

that variations will readily exist. For example, the depicted modules may be relocated from the client to server and vice versa (e.g., fulfillment may be performed at the server and the results returned to the client). Accordingly, the depicted placement of components and topology is merely one example of many possible configurations.

As discussed above, the depicted system may be used to address utterances which do not include proper entities (e.g., “Show me nearby restaurants”) as well as utterances which do include proper entities (e.g., “Tell me how to get to Guddu du Karahi”). A user **805** may speak a command **810** to a user interface **820** of a client device **815**. For example, the user may ask “Show me nearby restaurants”. The client device **815** may be an iPhone®, iPad®, tablet, personal computer, personal digital assistant, etc., or any device able to receive audio from the user **805**. The user interface **820** may convert the incoming command to a waveform **825a**. The waveform **825a** may be stored locally before being transmitted to the server **850**. Storing the waveform locally may allow portions of the waveform to later be considered by the client device, based on the hypotheses, without requesting that the waveform **825b** be transmitted back to the client from the server (one will recognize that in some other embodiments the server may instead transmit all or a portion of the waveform back to the client).

The server **850** may submit the waveform **825b** to a primary recognizer **830**. Primary recognizer **830** may be an “open-dictation” ASR system as known in the art. For example, the primary recognizer **830** may employ a lexicon associating energy patterns in a waveform with phonetic components to identify words corresponding to the phonetic components. Bayesian techniques as known in the art may be applied.

The server system **850** may include a Natural Language Understanding module (NLU) **855** configured to convert the transcription and word timings from the primary recognizer **830** into hypotheses. The hypotheses **815** and associated metadata may then be transmitted across a medium (e.g., the Internet) to client system **815**. Note, as discussed above, that where no proper entities appear in the utterance, a hypothesis may be generated without any acoustic spans. Where proper entities do appear (e.g., “Guddu du Karahi”) in the utterance, however, one or more acoustic spans may be indicated in the hypotheses. The hypotheses metadata may include the results of the ASR, such as the timestamps for word occurrences and the confidence of recognition for a given word.

The hypotheses may be received at a secondary recognizer **860**. The secondary recognizer **860** may be a grammar based ASR as discussed herein. If the hypotheses do not include acoustic spans, the hypotheses may pass through the scoring module **885** (if necessary) to identify a best match and proceed to the fulfillment unit **890**, possibly as a symbolic representation, which will attempt to fulfill the request (e.g., make a request to Yelp®). For the example request “Show me nearby restaurants”, e.g., the fulfillment unit **890** may contact a map server and request a list of restaurants within proximity to the user’s **805** coordinates. Once the results **870** have been retrieved, the client device may present the results to the user.

However, if the hypotheses do include acoustic spans (e.g., in the request “Tell me how to get to Guddu du Karahi”), the secondary recognizer **860** may consult adaptation object generator **865** to identify an appropriate grammar of proper names from the various components **840a-e** of the user device. Having identified possible proper names for the acoustic spans, the secondary recognizer **860** may apply

various of the identified proper entities and determine corresponding confidence levels. These decodings may be referred to the scoring module **885** so that the most likely candidate may be identified. The most likely candidate may then be passed to the fulfillment module **890** as discussed above.

Thus, inquiries which do not reference proper names of persons, locations, or the fanciful trademark names of different entities may be processed in a conventional manner. However, inquiries including such proper entities may also be readily identified using the secondary recognizer as described herein. Again, although FIG. **8** depicts a client-server architecture wherein the proper name resolution modules are split between the devices one will readily recognize variations upon this topology. For example, rather than a client-server architecture, all the discussed operations may occur on a single machine (e.g., the client device **815**).

Example Process Flow Overview

FIG. **9** is a flow diagram depicting the proper name recognition process at a high level for various embodiments using automatic speech recognition (ASR) and natural language understanding (NLU) components. As discussed herein, recognition of a proper name within an utterance may generally proceed in four steps, depicted in FIG. **9**. These steps are generally referred to herein as “Primary Recognition” **905**, “Understanding” **910**, “Secondary Recognition” **915**, and “Fusion” **920**. The client may perform the object preparation stage of “Adaptation Object Generation” at block **900a**, though one will recognize that the depicted order is merely for exemplary purposes and the process block may occur at other times in other embodiments than as depicted here.

“Primary Recognition” **905** for the example phrase “Tell me how to get to Guddu du Karahi” may occur at the open dictation (ASR) unit **830**. The output **805** of “Primary Recognition” **905** may include a nominal transcription, the start and end time within the waveform of each transcribed word at the granularity of a single frame, plus typically one or more putative proper name entity acoustic spans and information relating to the type of each span.

The NLU **855** may then perform the “Understanding” **910** step, applying the methods of natural language understanding to hypothesize one or more types to each span, possibly also adjusting the span boundaries (start frame and end frame) as assigned by the primary recognition step, and may provide additional information, such as potential shim words, and prefix and suffix acoustic context words, all defined in the sequel, that may aid in the decoding of each span.

Thus, the server may infer the presence of proper names in the text as described below and prepare one or more hypotheses **815** for their resolution. The hypotheses **815** may be submitted to the client. The client may then identify proper entities from the various components **840a-e** of the user device. For example, a GPS **840a** component may provide relevant street names near the user’s location, an address book **840c** may store the user’s **805** contacts, a search cache **840d** may reflect recent inquiries and operations performed by the user **805**, and a calendar **840b** may reflect meetings and events associated with user **805**. The content from one or more of these components may be considered when identifying proper entities as discussed herein.

They client may perform the finalization stage of “Adaptation Object Generation” at block **900b**. For example, the client may consult various local modules (e.g., the search cache) to identify appropriate proper entities to consider in

the grammar for “Secondary Recognition” **915**. One will recognize that the steps need not necessarily proceed in this order and that “Adaptation Object Generation” may occur earlier in the process.

The client may then perform “Secondary Recognition” **915**, by seeking to substitute various proper name entities for the acoustic spans to achieve suitable local resolution results. For example, the client may use an ASR or a separate grammar-based ASR system to determine the probability that a given portion of the waveform corresponds to a proper name identified from the components **840a-d**.

During “Secondary Recognition” **915**, each such span may now be decoded by a grammar-based speech recognizer within client-side proper name resolutions, using a grammar that has been specially adapted to the type and individual user of the system based upon components **840a-d**. Whereas the adaptation of an open dictation recognizer to a specialized vocabulary or context may be computationally expensive, a new grammar, or a grammar with unpopulated placeholder “slots,” can be generated and ready for service on the client in a few seconds or less.

The output of the “Secondary Recognition” **915** step, performed on the subject acoustic span, using a suitably specialized grammar, may be considered the nominal transcription of the span. The literal sequences of this same grammar may be labeled with the meaning of each potential decoding path through the grammar. Thus the generation of the transcription can at the same time generate an appropriate symbolic meaning, for the selected decoding path. If only the portion of the original utterance associated with the proper name is provided as the audio input to a grammar-based ASR system, and the grammar used for decoding comprises all and only the business names extracted from the user’s calendar for that day, then it is highly likely that the correct proper name will be decoded.

If no competing alternative hypotheses regarding the presence, type or extent of each span, were proposed by the prior processing steps, then the “Secondary Recognition” **915** may be complete and the decoded result may be submitted for fulfillment. However, this is sometimes not the case where several alternative decodings may have been hypothesized (which may or may not include acoustic spans corresponding to proper entities), each with associated NLU and ASR confidence scores. In this situation the client module may select the final preferred decoding, or assign a confidence score to each whole decoding, and provide a ranked list of alternatives to the server. As differing hypotheses may comprise different numbers of acoustic spans, this may force the comparison of hypotheses that are based upon different numbers of confidence scores. The “Fusion” **920** of the different scores may occur when the proper names are considered in the context of the NLU unit.

As discussed above, the “Primary Recognition” step **905**, may be accomplished by the open dictation ASR **830** technology, which may be more powerful and more flexible, but more computationally demanding than a grammar-based ASR. The “Primary Recognition” step **905** may not bear the principal responsibility for generating the final transcription of the input utterance in some embodiments recognizing proper names. Rather, this step may determine the portion or portions of the input waveform that comprise one or more of the proper name entities. For example, the open dictation ASR **830** may simply note portions of the waveforms for which the identified words have exceedingly low confidence levels.

As an open dictation ASR **830** may be used initially, this approach may achieve high accuracy recognition and under-

23

standing of proper names and similar entities (hereafter “proper name entities”), occurring within utterances that as a whole are not constrained to conform to an ASR grammar. Moreover the proper name entities may contain or be comprised wholly of words not present in the vocabularies of the ASR systems as normally constituted.

The same technique may be used to recognize entities, such as numbered street addresses, that include within them street names and possibly city and state names as well. An example of the latter would be “333 Ravenswood Avenue” or the more precise “333 Ravenswood Avenue, Menlo Park, Calif.” Finally, word sequences that are purely descriptive and generic, such as “grandma’s house,” “the office,” “day-care,” “the country club” and so on, which the user has identified to the system as personally significant, can also be treated by the process of FIG. 9.

Some embodiments place additional sources of information—those derived from the intermediate NLU processing step—at the disposal of the speech decoding and meaning assignment process as a whole. In some embodiments, the open dictation ASR component of the system, prepared by the methods described here, requires no further adaptation or modification to enable recognition of names and entities that are not even present in its vocabulary. Thus, this component may be shared by a multitude of users, with the necessary adaptations to enable recognition of proper names confined to other components of the system.

Generalized Proper Name Recognition Process

FIG. 10 is a flow diagram depicting various steps in a proper name recognition process as may occur in some embodiments. At block 1005, the system may receive an utterance waveform from a user. Where the process is divided between client and server devices a copy of the waveform may be retained at the client as discussed herein.

At block 1010, a “standard” open dictation ASR may be applied to the waveform. This may produce a complete textual word for every aspect of the waveform, even when the confidence levels are exceptionally low. However, some embodiments further contemplate applying a modified version of the open dictation ASR to the waveform to achieve one or more textual readings that explicitly identify words that may reflect proper names (e.g., based on the highest possible confidence level for a word still failing to exceed a threshold). These modified systems may indicate placeholder words for the potential proper names (e.g., fna, lna, and sa designations as discussed herein). Block 1010 may roughly correspond to the “Primary Recognition” step 905. Block 1020 may roughly correspond to the “Understanding” step 910.

At block 1030, the system may determine if one or more word confidence values are deficient, e.g., have confidence levels falling below a threshold, or if the modified systems have otherwise identified one or more potential proper names. Where all of the confidence values exceed a threshold, or where no proper name candidates are otherwise identified, the system may transition to block 1035. At block 1035, the system may complete processing to generate a symbolic representation of the request. At block 1040, the system may attempt fulfillment using the symbolic representation and return any results to the user. As discussed above, one will recognize that fulfillment is just one possible application for the above processes. Accordingly, blocks 1035 and 1040 may readily be substituted by other applications, e.g., performing operations on the client device.

Conversely, if possible proper names are identified at block 1030, then at block 1045 the system may generate one or more hypotheses based upon the deficient word(s) that

24

include acoustic spans as described in greater detail herein. At block 1050, the system, e.g., the client device, may decode each probable first name segment against its first name grammar. Block 1050 may generally correspond to the “Secondary Recognition” 915 step. In the some embodiments, the “Secondary Recognition” 915 step reduces to little more than inserting the most likely grammar decoding result in the appropriate location in the text output by “Primary Recognition” 905 and/or “Understanding” 910 operations.

At block 1055, the system may determine which of the proposed proper entities for the acoustic spans (and/or the confidence levels associated with a hypothesis without acoustic spans) best corresponds to the utterance. For example, the system may identify the resolution with the highest cumulative confidence values. This determination may be made by considering one or more of the original, open dictation ASR confidence values, the original NLU confidence values, the ASR grammar-based confidence values determined at block 1050, and possibly a second NLU determination using the ASR grammar-based results, as part of a “Score Fusion” 920.

If an appropriate resolution is identified, the system may convert the proper name to symbolic form at block 1060 and present the symbolic representation of the entire utterance for fulfillment. Conversely, if no appropriate resolutions are found at block 1055, the system may announce a failure at block 1065. In some embodiments, rather than announce failure, the system may instead attempt fulfillment with the words having deficient probabilities or with the closest approximates.

“Understanding”/Hypothesis Creation Example

FIG. 11 is an example hypothesis corpus as may be generated in some embodiments. These example hypotheses may be generated as part of block 1045. A waveform 1105 may be associated with the user utterance “Where is Guddu de Karahi, the restaurant, located?” 1110. The client ASR/NLU and/or the server ASR/NLU may generate the proposed decodings 1115a-c. The decoding 1115a construes the utterance as “Where is goose karate the restaurant, located?” with confidence values of 110 and 150 associated with the words “goose” and “karate” respectively. This example, where low confidence values are generated, but words are identified anyway, may correspond to the path through blocks 1015 and block 1045 discussed above. These confidence values may be lower than a threshold, e.g., 300, indicating an incorrect association. In this example, “goose” mismatches “Guddu” and “karate” mismatches “Karahi” as the words are superficially similar. Accordingly, the 110 and 150 confidence levels reflect an unlikely match (e.g., because the spectral character of the waveform doesn’t agree with the expected character of the phonemes in these words). However, if no better proper name match is found for the proposals 1115b-c, the system may accept this interpretation by default, and submit these words to the symbolic representation for fulfillment.

With regard to the second proposed decoding 1115b, the client ASR/NLU and/or the server ASR/NLU may construe “Guddu” as “parking for” with corresponding low confidence levels 90 and 75. Here, the system may have simply identified the portion of the waveform within “Karahi” as unknowable and accordingly, a potential proper name. As discussed below, an appropriate substituted identifier (fna, lna, etc.) may be inserted for the hypothesis.

Finally, in proposed decoding 1115c, the system may simply have recognized the entirety of the “Guddu de Karahi” waveform as being unrecognizable. The system

25

may recognize that two separate words were spoken, but may be unable to recognize the identity of the words. The examples of decodings **1115b** and **1115c**, where placeholders are used to identify possible proper names, may correspond to the path through blocks **1010** and block **1020** discussed above.

FIG. **12** is an example of a first hypothesis breakdown based upon the example of FIG. **11** as may occur in some embodiments. With regard to the second proposal **1115b**, the system may recognize that the “Karahi” portion between 1520 ms and 1750 ms could not be recognized. Accordingly, a hypothesis **1205** having an acoustic span between 1520 and 1750 may be generated (one will recognize that the values 1520 and 1750 are merely exemplary and other representations, e.g., milliseconds, may be used). The NLU may infer that this is a “Location inquiry” and ascribe a corresponding potential meaning, based upon the “Where” and “parking for” portions of the utterance. Similarly, the NLU may infer based upon the phrase “restaurant” that the span is of type “Business name”. The Potential Meaning and Putative Type in the hypothesis may be used to localize the search for proper names on the client device. For example, knowing that this is a “Location Query” the client device may not consider first and last names in an address book, but may rather consider only meeting locations in a calendar. Only business names associated with locations in the calendar may be considered based upon the Putative Type of span.

As discussed herein, the pronunciation of a proper noun may be influenced by the preceding and succeeding words. Accordingly, a prefix portion and a suffix portion may also be identified in the hypothesis for consideration by the components searching for proper names. Alternatively, some embodiments may prepend and post-pend brief segments of silence (or low-power background noise), ramped from very low power to the nominal power of the utterance (e.g., ramping up from low to nominal power, for the prepended audio, and ramping down from nominal power to low for the post-pended audio). This temporal smoothing of the audio input may eliminate abrupt audio transitions, which could be falsely matched as fricative phonemes.

If the words adjacent to the placeholder are decoded with notably low confidence scores—or if an initial decoding of a given audio segment by the secondary recognizer yields an anomalously low confidence score—some embodiments perturb the nominal start and end times of the extracted audio segment, thereby producing multiple candidate segments for decoding. All of these may then be passed as variants to “Secondary Recognition” **915**, which can decode them all and select the decoding with the highest confidence score as the nominal answer.

FIG. **13** is an example of a second hypothesis breakdown based upon the third proposal **1115c** in the example of FIG. **11** as may occur in some embodiments. With regard to the third proposal **1115c**, the system may recognize that the “Guddu de Karahi” portion between 1220 ms and 1750 ms could not be recognized. Accordingly, a hypothesis **1205** having an acoustic span between 1220 ms and 1750 ms may be generated, and the prefix and suffix portions adjusted accordingly. The NLU may again identify the proper noun as a “Business Name” in the Putative Type but may instead consider the general inquiry as an “Address Book” query, limiting the search to only the address book contents. “Understanding”/Hypothesis Generation Process

FIG. **14** is a flow diagram depicting various steps in a server-side process for proper name recognition as may occur in some embodiments. For example, process **1400**

26

may depict the operations of block **1030** in greater detail. At block **1405**, the system may consider the next possible textual representation generated from the ASR and/or NLU. A plurality of probabilities and word timings may be included as part of the textual representation.

At block **1410**, the system may prepare a hypothesis template, e.g., a data structure for holding the various hypothesis parameters.

At block **1415**, the system may generate a “potential meaning” for the hypothesis by referencing NLU statistics.

At block **1420**, the system may generate a “putative type” for the span by referencing NLU statistics.

At block **1430**, the system may determine the timestamps associated with the beginning and end of the span. As discussed in greater detail below, the prefix and suffix to the potential proper name in question may also be included in this determination.

At block **1435**, the system may consider additional potential text representations if they exist. If not, the system may proceed to block **1440**, where the system may submit the queued hypotheses to the client system for analysis, or depending upon the topology, to the appropriate component for analyzing the hypotheses. For example, in some embodiments, the system may analyze the hypotheses locally on the server, or they may be both generated and analyzed on the client device.

“Secondary Recognition”—Client-Side Process

FIG. **15** is a flow diagram depicting various steps in a client-side process **1500** for proper name recognition as may occur in some embodiments. At block **1505** the client module may consider the next hypothesis received from the server module.

At block **1510**, the client module may extract the potential meaning from the hypothesis. At block **1515**, the client module may extract the putative type of span from the hypothesis. At block **1520**, the client module may collect the corpus of proper nouns based upon the potential meaning and/or putative type.

At block **1525**, the system may extract the timestamps associated with the putative span and (if present) the timestamps to any suffix or prefix portions. At block **1530**, the system may consider the next proper name in the identified corpus. Where substitution of the corpus member results in a satisfactory confidence values at block **1535**, the system may include the substituted member among the successful resolutions at block **1540**.

If not all proper names in the corpus have been considered at block **1545**, the system may continue with the next corpus member.

At block **1550**, the system determines which resolution to submit for fulfillment, e.g., using the score fusion processes discussed herein.

“Score Fusion”

As discussed above, it is likely that more than one hypothesis and more than one solution to a hypothesis have been generated following “Secondary Recognition”. Furthermore, as in the example hypotheses from decodings 2 and 3, differing hypotheses may comprise different numbers of acoustic spans. This may require that hypotheses based upon different numbers of confidence values be considered so as to achieve a meaningful and reliable score-based ranking. In some embodiments, the ranking of hypotheses may be made purely upon these ASR confidence scores. However, various embodiments contemplate including the NLU system **855** to influence this ranking. If the NLU confidence scores can be normalized to probabilities, they

27

may be meaningfully combined with the grammar and/or open dictation ASR confidence scores.

The following is one example for performing this combination. Let T_i and A denote the transcription (e.g., text **880**) and acoustic input (e.g., waveform **825**), and let M_i denote the symbolic meaning assigned by NLU processing to the i th hypothesis. The expression $P(M_i|T_i)$ reflects the NLU confidence score of the i th hypothesis meaning, given the associated transcription.

By the product law for conditional probabilities:

$$P(M_i, T_i|A) = P(M_i|T_i, A)P(T_i|A)$$

Assuming the approximation:

$$P(M_i, T_i|A) \approx P(M_i|T_i)$$

This then yields

$$P(M_i, T_i|A) \approx P(M_i|T_i)P(T_i|A)$$

which expresses the confidence that the correct meaning and transcription of the (fixed) acoustic input A have been acquired.

Grammar Preparation

Following "Primary Recognition" **905**, the system has identified not only the start and end time of each such segment, but the likely type of the name in question—that is, a person's first name, a person's last name, a street name, and so on. A specialized grammar may be used for each such name type. These grammars may be relatively small compared to the full space of names of the appropriate type, but which nevertheless have a high probability of containing the name that was spoken.

Each individual name type grammar may be prepared from an appropriate data source, specialized to information about the user's friends and associates, location, past, current or future activities, and so on. For instance, a first name grammar may be prepared by listing all the first names of any contact found in the user's address book, along with common nicknames or abbreviations; similarly with last names. Likewise a street name grammar may be prepared by combining the names of all streets within a given radius of the user's current location, possibly augmented by all street names extracted from past or future appointments, as noted in the user's personal calendar, or all streets on or near any recently-driven routes, as determined by a car or telephone handset GPS system. As discussed above, one useful characteristic of this architecture is that these grammars may be prepared at the client, and never communicated to the server.

Multi-Word Sequence Aggregates

It may happen that some names to be incorporated in an aggregate may comprise not just a single word, but a sequence of words. This arises naturally in the consideration of street names, which are typically composed of both a name and type, the latter of which may or may not be verbalized. Consider for instance "Rengstorff Avenue," "Fourteenth Street," and "Riverside Drive", each of which may also be spoken respectively as "Rengstorff", "Fourteenth" or "Riverside". Note that "avenue" is not infrequently spoken as the single syllable "ave," rhyming with "have." This suggests that whatever solution is chosen for handling the street type, two variant forms for the "avenue" type may be required. Moreover, regardless of the notion of street type, some proper names have multiple words, for example "The Embarcadero," "Dry Creek [Road]," or "El Camino Real."

For these latter cases, various embodiments concatenate together those elements of the name's word sequence that are obligatory, in the sense that they will always be verbal-

28

ized. The resulting object may be treated as a single word, and include its pronunciation may be included in the aggregate. Thus some embodiments may treat each of "The_Embarcadero," "Dry_Creek," and "El_Camino_Real" as a single word.

Two approaches are contemplated for words in such sequences that might not be verbalized. The first approach is to concatenate together all the words associated to a given name as just explained, both with and without the optional words. All variants would be included in the aggregate. Thus considering the cases exhibited above would produce: Rengstorff_Avenue; Rengstorff_Ave; Rengstorff; Fourteenth_Street; Fourteenth; Riverside_Drive; Riverside; Dry_Creek; Dry_Creek_Road. This will cover the acoustic space well.

The second approach is to introduce an additional aggregate to capture the type, say street-type-aggregate or sta, and include within it the pronunciations of all the types. This may approximately halve the number of pronunciations nominally included in sa. However it may weaken the language model, and thereby hamper the ability of the primary recognizer to find the end of the audio segment that comprises the street name.

Alternative Method for Language Model Generation

The method described above for language model generation produces reasonable values for each aggregate, in quite general contexts. However it may be time-consuming and its full generality may not be required.

Accordingly, some embodiments implement a more restricted method, which may yield good results in the contexts in which the technique is likely to be the most useful. This method may preprocess the entire training corpus with the NLU system, replacing proper name entities with appropriate aggregate words, in context. This will then yield n-gram counts with aggregate words, from which language models can be constructed, e.g., by conventional means, with such words as first-class objects.

Some embodiments may adopt a hybrid approach in which the conditional probabilities $p(\text{fna-i}|\text{lh})$ could be determined by this method while values for $p(\text{x}|\text{lh})$, with $\text{fna-i} \rightarrow \text{fna-i} \rightarrow \text{fna-i}$, could be determined by the previously outlined method.

Intentional Non-Match Grammar to Proper Name Type

The "Primary Recognition" **905** and "Understanding" **910** steps may return the sequence $\text{fna-i} \text{ lna-j}$. The method proposed above would perform a "Secondary Recognition" **915** decoding of the audio segment associated to fna-i with a grammar of first names, and an independent "Secondary Recognition" **915** decoding of the audio segment associated to lna-j with a grammar of last names.

However, these two independent decodings could conceivably yield a name that does not correspond to an individual appearing in the user's address book. The sequence $\text{fna-i} \text{ lna-j}$ in the output of the primary decoder may correspond to a single whole name. Thus the entire stretch of audio, from the start of the fna segment to the end of the lna segment, may be recognized during "Secondary Recognition" **915** against a grammar that consists of all full contact names, as they appear in the user's address book.

Semantic Labeling of Placeholder Decodings

In some embodiments, e.g., where the command includes a request to send a message to a recipient, it may be necessary to extract, as part of the meaning of the utterance, the intended recipient of a message. For example, consider the utterance "send a message to Barack thanks so much for the invitation comma we'd love to visit you and Michelle the next time we're in Washington". The output of the primary recognizer may then very well read:

send a message to fna thanks so much for the invitation comma we'd love to visit you and fna the next time we're in Washington

Note that two audio segments are identified here as likely first names and both may be marked for decoding by the "Secondary Recognition" 915 recognizer (e.g., the grammar-based ASR). Assuming that "Barack" and "Michelle" are both in the user's contact list, these decodings will probably contain the correct results. "Barack" can then be mapped to a suitable phone number or email address to use as the destination of the message.

However, as discussed herein, the NLU may be located at the server rather than the client. Without further communication from the client back to the server, of the "Secondary Recognition" 915 recognizer results, there may be no way to perform the required analysis to determine that "Barack" is indeed the name of the intended recipient.

Accordingly, in some embodiments the NLU will be able to work out the position of the intended recipient, from the information that a particular token (fna) in the decoding is likely to be a person's proper name, and from the words that appear adjacent or near to this token. This information may be communicated to the client, where the "Secondary Recognition" 915 recognizer can definitively identify the recipient name. Other elements of the client software may process this name to determine a suitable destination address.

One way to communicate this information from the NLU to the "Secondary Recognition" 915 recognizer would be to add an appropriate field, comprising symbolic meaning, to the protocol element that identifies the audio segment as an object for processing by the secondary decoder. (In cases where the segment has no special meaning, as in the second instance of fna in the example above, this field may contain "null" or some other neutral value.)

In some embodiments, the putative type of the audio segment could be changed, in a manner understandable to the client, to communicate to the client both the grammar to be used by the secondary decoder, and the special meaning, if any, of the audio segment. For example, the first instance of fna could be changed to a type fna-recipient, with the client suitably modified to decode the associated audio segment against the first name grammar as before, and then interpret the result as the name of the intended recipient.

Grammars Designed for Span Extent Errors

Proper selection of the span extent may improve secondary decoding. Furthermore, some embodiments allow for coarticulation effects in selecting phone models during the decoding process. These issues may be dealt with by expanding the acoustic span to include some number of acoustic prefix words and acoustic suffix words, which are those immediately preceding and following the nominal proper name entity. This yields the important distinction between the target span, which is the span of words comprising the nominal proper name entity, and the full span, which includes the audio putatively corresponding to the just-mentioned acoustic prefix words and acoustic suffix words. Moreover the secondary recognition grammar must then be structured in such a way that allows decoding of these words. Indeed, it may be helpful if the secondary recognition proceeds through the first acoustic prefix word and the final acoustic suffix word.

This may make the indicated secondary recognition less sensitive to the nominal start and end of the full span, and the secondary recognition can then choose freely just where the proper name entity itself begins and ends within the full span. Moreover, by properly structuring the grammar, it can be arranged that the audio corresponding to the prefix words

immediately preceding the putative proper name entity words, and likewise the suffix words immediately following, can be absorbed into well-matching words within the active grammar. This may allow a word decoded as "to" by the primary recognizer, and therefore marked as external to the target span, may in fact be the onset of the proper name "Toby," which is present among the names in the active grammar. By making the decoding of this acoustic prefix word optional (e.g., by providing an epsilon-path around it), the appropriate frames of audio may thereby participate in the successful decoding of the name "Toby."

It may be desirable to structure the grammar associated to the prefix and suffix words so that those words may themselves be subdivided at plausible acoustic boundaries. For example, if the immediately preceding acoustic prefix word is "filbert," the grammar that embeds this should allow the word to be phonetically divided within the decoding process into "fill" and "bert", thereby enabling the audio associated with the latter to be absorbed in decoding the proper name "Bert," should it happen to have been spoken, and present within the active grammar. This may be accomplished by comparing the nominal phoneme sequence of the primary decoding with the contents of the vocabulary, and using the language model to hypothesize plausible alternate word divisions, which can then be reflected in the associated grammar structure.

While various of the foregoing aspects may be appropriate to span-too-small errors, the inverse problem can occur as well. The system may generate a span-too-large error, and incorrectly guesses that audio corresponding to a word or words that is in fact external to the proper name entity lies within it.

To deal with this problem, some embodiments contemplate "shim words". Shim words are words that should be present in the primary recognizer decoding, adjacent to the target span, but for which the audio has erroneously been incorporated into the target span. In order to allow the secondary recognition to operate properly and suppress the matching of the associated frames against target words in the grammar, the grammar may be enlarged with optional paths that include such shim words.

As there may be no evidence that shim words are necessary, or what the shim words should be, they may be hypothesized using a forward (conventional) language model that identifies likely forward extension of the acoustic prefix words. A backward language model that identifies likely backward extension of the acoustic suffix words may similarly be used. These considerations may therefore yield one or more such words, which may be incorporated as optional alternatives within the target grammar shim words (and selection thereof via language models).

Computer System

FIG. 18 is a block diagram of a computer system as may be used to implement features of some of the embodiments. The computing system 1800 may include one or more central processing units ("processors") 1805, memory 1810, input/output devices 1825 (e.g., keyboard and pointing devices, display devices), storage devices 1820 (e.g., disk drives), and network adapters 1830 (e.g., network interfaces) that are connected to an interconnect 1815. The interconnect 1815 is illustrated as an abstraction that represents any one or more separate physical buses, point to point connections, or both connected by appropriate bridges, adapters, or controllers. The interconnect 1815, therefore, may include, for example, a system bus, a Peripheral Component Interconnect (PCI) bus or PCI-Express bus, a HyperTransport or industry standard architecture (ISA) bus, a small computer

system interface (SCSI) bus, a universal serial bus (USB), IIC (I2C) bus, or an Institute of Electrical and Electronics Engineers (IEEE) standard 1394 bus, also called "Firewire".

The memory **1810** and storage devices **1820** are computer-readable storage media that may store instructions that implement at least portions of the various embodiments. In addition, the data structures and message structures may be stored or transmitted via a data transmission medium, such as a signal on a communications link. Various communications links may be used, such as the Internet, a local area network, a wide area network, or a point-to-point dial-up connection. Thus, computer readable media can include computer-readable storage media (e.g., "non transitory" media) and computer-readable transmission media.

The instructions stored in memory **1810** can be implemented as software and/or firmware to program the processor(s) **1805** to carry out actions described above. In some embodiments, such software or firmware may be initially provided to the processing system **1800** by downloading it from a remote system through the computing system **1800** (e.g., via network adapter **1830**).

The various embodiments introduced herein can be implemented by, for example, programmable circuitry (e.g., one or more microprocessors) programmed with software and/or firmware, or entirely in special-purpose hardwired (non-programmable) circuitry, or in a combination of such forms. Special-purpose hardwired circuitry may be in the form of, for example, one or more ASICs, PLDs, FPGAs, etc.

Remarks

The above description and drawings are illustrative and are not to be construed as limiting. Numerous specific details are described to provide a thorough understanding of the disclosure. However, in certain instances, well-known details are not described in order to avoid obscuring the description. Further, various modifications may be made without deviating from the scope of the embodiments. Accordingly, the embodiments are not limited except as by the appended claims.

Reference in this specification to "one embodiment" or "an embodiment" means that a particular feature, structure, or characteristic described in connection with the embodiment is included in at least one embodiment of the disclosure. The appearances of the phrase "in one embodiment" in various places in the specification are not necessarily all referring to the same embodiment, nor are separate or alternative embodiments mutually exclusive of other embodiments. Moreover, various features are described which may be exhibited by some embodiments and not by others. Similarly, various requirements are described which may be requirements for some embodiments but not for other embodiments.

The terms used in this specification generally have their ordinary meanings in the art, within the context of the disclosure, and in the specific context where each term is used. Certain terms that are used to describe the disclosure are discussed below, or elsewhere in the specification, to provide additional guidance to the practitioner regarding the description of the disclosure. For convenience, certain terms may be highlighted, for example using italics and/or quotation marks. The use of highlighting has no influence on the scope and meaning of a term; the scope and meaning of a term is the same, in the same context, whether or not it is highlighted. It will be appreciated that the same thing can be said in more than one way. One will recognize that "memory" is one form of a "storage" and that the terms may on occasion be used interchangeably.

Consequently, alternative language and synonyms may be used for any one or more of the terms discussed herein, nor is any special significance to be placed upon whether or not a term is elaborated or discussed herein. Synonyms for certain terms are provided. A recital of one or more synonyms does not exclude the use of other synonyms. The use of examples anywhere in this specification including examples of any term discussed herein is illustrative only, and is not intended to further limit the scope and meaning of the disclosure or of any exemplified term. Likewise, the disclosure is not limited to various embodiments given in this specification.

Without intent to further limit the scope of the disclosure, examples of instruments, apparatus, methods and their related results according to the embodiments of the present disclosure are given below. Note that titles or subtitles may be used in the examples for convenience of a reader, which in no way should limit the scope of the disclosure. Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this disclosure pertains. In the case of conflict, the present document, including definitions will control.

What is claimed is:

1. A computer-implemented method for recognizing and understanding spoken commands that include one or more proper name entities, comprising:

receiving an utterance from a user;

performing primary automatic speech recognition (ASR) processing upon said utterance with a primary automatic speech recognizer to output a dataset comprising at least a sequence of nominal transcribed words and putative start and end times for each nominal transcribed word within said utterance;

performing understanding processing upon said dataset with a natural language understanding (NLU) processor to generate and augment the dataset with a nominal meaning for the utterance and to determine putative presence and type of one or more spoken proper name entities within said utterance, wherein a contiguous section of audio within said utterance corresponding to each putative proper name entity, as determined from said start and end times of the words of the putative proper name entity as transcribed by the primary automatic speech recognizer, comprises an acoustic span;

performing secondary automatic speech recognition (ASR) processing upon each said acoustic span with a secondary automatic speech recognizer, in each instance said secondary automatic speech recognizer specialized to process a given putative type of acoustic span to generate a nominal correct transcription and associated meaning for each said acoustic span;

substituting the nominal correct transcription and associated meaning obtained from each secondary recognition as appropriate within the dataset to revise the results of the primary automatic speech recognizer and natural language understanding processor; and outputting a complete and accurate transcription and meaning for the entire utterance.

2. The method of claim 1, further comprising:

specializing the secondary ASR recognizer by using an adaptation grammar comprising structure and content appropriate to a putative span type, as determined by NLU processing.

33

3. The method of claim 2, further comprising:
 deriving proper name entities defining structure and content of the adaptation grammar used to specialize the secondary ASR recognizer from any of:

- (a) business names resulting from a search;
- (b) business names, personal contact names, or both retrieved from any of a personal phone book, a personal calendar, or both;
- (c) contents of a music library or personal music storage device, including artist names, song names, album names, and genre names;
- (d) contents of a video library or personal video storage device, including actor names, director names, and genre names;
- (e) word sequences that the user has identified as personally significant; and
- (f) any combination of the foregoing.

4. The method of claim 2, further comprising:
 associating one or more meaning variables and decoding-dependent value assignments with pathways within said adaptation grammar with appropriate values thereof reported as part of the secondary ASR recognizer output to provide symbolic representation of an identity of a specific proper name entity recognized within the user's utterance.

5. The method of claim 2, further comprising:
 deriving proper name entities defining structure and content of the adaptation grammar used to specialize the secondary ASR recognizer from any of:

- (a) numbered street addresses associated to a particular locale;
- (b) roadway intersections associated to a particular locale;
- (c) landmarks associated to a particular locale; and
- (d) any combination of the foregoing; and

wherein said locale is determined by analysis of the utterance received from the user and supplied as an element of the span type.

6. The method of claim 2, further comprising:
 deriving proper name entities defining structure and content of the adaptation grammar used to specialize the secondary ASR recognizer from any of:

- (a) numbered street addresses associated to a particular locale;
- (b) roadway intersections associated to a particular locale;
- (c) landmarks associated to a particular locale; and
- (d) any combination of the foregoing; and

wherein said locale is determined by analysis of the physical location of the user and supplied as an element of the span type.

7. The method of claim 1, further comprising:
 specializing the secondary ASR recognizer by using an adaptation grammar of structure and content appropriate to a putative span type, as determined by NLU processing, said adaptation grammar additionally including acoustic prefix words, acoustic suffix words, or both, as transcribed by the primary ASR recognizer to ensure high accuracy secondary ASR recognition in view of coarticulation effects in the processed utterance and potential imprecise determination of span start and end times; and

correspondingly expanding said span to include said acoustic prefix words, acoustic suffix words, or both.

8. The method of claim 7, further comprising:
 including acoustic prefix words, acoustic suffix words, or both within the adaptation grammar by preparing said adaptation grammar as a slotted grammar with appropriate one or more prefix slots, suffix slots, or both and

34

populating said slots as appropriate with acoustic prefix words, acoustic suffix words, or both, as transcribed by the primary ASR recognizer.

9. The method of claim 7, further comprising:
 deriving the proper name entities defining structure and content of the adaptation grammar used to specialize the secondary ASR recognizer from any of:

- (a) business names resulting from a search;
- (b) business names, personal contact names or both retrieved from any of a personal phone book, a personal calendar, or both;
- (c) contents of a music library, including artist names, song names, album names, and genre names;
- (d) contents of a video library, including actor names, director names, and genre names;
- (e) word sequences that the user has identified as personally significant; and
- (f) any combination of the foregoing.

10. The method of claim 7, further comprising:
 associating one or more meaning variables and decoding-dependent value assignments with pathways within said adaptation grammar with appropriate values thereof reported as part of the secondary ASR recognizer output provide symbolic representation of an identity of a specific proper name entity recognized within the user's utterance.

11. The method of claim 7, wherein the proper name entities defining the structure and content of the adaptation grammar used to specialize the secondary ASR recognizer are derived from any of:

- (a) numbered street addresses associated to a particular locale;
- (b) roadway intersections associated to a particular locale;
- (c) landmarks associated to a particular locale; and
- (d) any combination of the foregoing; and

wherein said locale is determined by analysis of the utterance received from the user and supplied as an element of the span type.

12. The method of claim 7, further comprising:
 deriving the proper name entities defining structure and content of the adaptation grammar used to specialize the secondary ASR recognizer from any of:

- (a) numbered street addresses associated to a particular locale;
- (b) roadway intersections associated to a particular locale;
- (c) landmarks associated to a particular locale; and
- (d) any combination of the foregoing; and

wherein said locale is determined by analysis of the physical location of the user and supplied as an element of the span type.

13. The method of claim 7, further comprising any of:

- (a) constructing said adaptation grammar to render recognition of acoustic prefix words and acoustic suffix words optional to decode, in whole or part, to tolerate span-too-small errors in determination of the target span;
- (b) constructing said adaptation grammar to include one or more optional-to-decode shim words that immediately follow acoustic prefix words, to include one or more optional-to-decode shim words that immediately precede acoustic suffix words, or both, to tolerate span-too-large errors in determination of the target span; and
- (c) constructing said adaptation grammar by simultaneously comprehending both (a) and (b).

35

14. The method of claim 13, further comprising any of:
 (a) using a forward language model that identifies likely forward extensions of any acoustic prefix word or words to determine one or more optional-to-decode shim words immediately following acoustic prefix words;

(b) using a backward language model that identifies likely backward extensions of any acoustic suffix word or words to determine one or more optional-to-decode shim words immediately preceding acoustic suffix words; and

(c) both of (a) and (b).

15. The method of claim 13, further comprising:

including acoustic prefix words, acoustic suffix words, optional-to-decode shim words immediately following acoustic prefix words, or optional-to-decode shim words immediately preceding acoustic suffix words, or any combination thereof within the adaptation grammar by preparing said grammar as a slotted grammar with appropriate one or more prefix slots, suffix slots, shim slots immediately following prefix slots, or shim slots immediately preceding suffix slots, or any combination thereof, and populating said prefix slots, suffix slots, or both, as appropriate with acoustic prefix words, acoustic suffix words, or both, as transcribed by the primary ASR recognizer, and populating said shim slots immediately following prefix slots with appropriate optional-to-decode shim words immediately following acoustic prefix words, or populating said shim slots immediately preceding suffix slots with appropriate optional-to-decode shim words immediately preceding acoustic suffix words, or both.

16. An apparatus for recognizing and understanding spoken commands that includes one or more proper name entities, comprising:

36

a processor configured for receiving an utterance from a user;

said processor performing primary automatic speech recognition (ASR) processing upon said utterance with a primary automatic speech recognizer to output a dataset comprising at least a sequence of nominal transcribed words and putative start and end times for each nominal transcribed word within said utterance;

said processor performing natural language understanding (NLU) processing upon said dataset with a natural language understanding processor to generate and augment the dataset with a nominal meaning for the utterance and to determine putative presence and type of one or more spoken proper name entities within said utterance, wherein a contiguous section of audio within said utterance corresponding to each putative proper name entity, as determined from said start and end times of the words of the putative proper name entity as transcribed by the primary automatic speech recognizer, comprises an acoustic span;

said processor performing secondary automatic speech recognition processing upon each said acoustic span with a secondary automatic speech recognizer, in each instance said secondary automatic speech recognizer specialized to process a given putative type of acoustic span to generate a nominal correct transcription and associated meaning for each said acoustic span;

said processor substituting the nominal correct transcription and associated meaning obtained from each secondary recognition as appropriate within the dataset to revise the results of the primary automatic speech recognizer and natural language understanding processor; and

said processor outputting a complete and accurate transcription and meaning for the entire utterance.

* * * * *